CLC **Protein** Workbench

User manual

Manual for

CLC Protein Workbench 5.8

Windows, Mac OS X and Linux

February 23, 2012

This software is for research purposes only.

CLC bio
Finlandsgade 10-12
DK-8200 Aarhus N
Denmark

Contents

I Introduction **9**

1 Introduction to *CLC Protein Workbench* **10**

1.1 Contact information . . . 12
1.2 Download and installation . . . 12
1.3 System requirements . . . 15
1.4 Licenses . . . 15
1.5 About CLC Workbenches . . . 27
1.6 When the program is installed: Getting started . . . 29
1.7 Plug-ins . . . 30
1.8 Network configuration . . . 33
1.9 The format of the user manual . . . 34

2 Tutorials **36**

2.1 Tutorial: Getting started . . . 37
2.2 Tutorial: View sequence . . . 38
2.3 Tutorial: Side Panel Settings . . . 40
2.4 Tutorial: GenBank search and download . . . 43
2.5 Tutorial: BLAST search . . . 44
2.6 Tutorial: Tips for specialized BLAST searches . . . 47
2.7 Tutorial: Proteolytic cleavage detection . . . 52
2.8 Tutorial: Align protein sequences . . . 54
2.9 Tutorial: Create and modify a phylogenetic tree . . . 56
2.10 Tutorial: Find restriction sites . . . 57

II Core Functionalities **60**

3 User interface **61**

3.1 Navigation Area . . . 62
3.2 View Area . . . 69
3.3 Zoom and selection in View Area . . . 76
3.4 Toolbox and Status Bar . . . 78
3.5 Workspace . . . 79
3.6 List of shortcuts . . . 80

4 Searching your data **83**

4.1 What kind of information can be searched? . . . 83
4.2 Quick search . . . 84
4.3 Advanced search . . . 86
4.4 Search index . . . 88

5 User preferences and settings **89**

5.1 General preferences . . . 89
5.2 Default view preferences . . . 90
5.3 Advanced preferences . . . 93
5.4 Export/import of preferences . . . 94
5.5 View settings for the Side Panel . . . 95

6 Printing **98**

6.1 Selecting which part of the view to print . . . 99
6.2 Page setup . . . 100
6.3 Print preview . . . 101

7 Import/export of data and graphics **102**

7.1 Bioinformatic data formats . . . 102
7.2 External files . . . 109
7.3 Export graphics to files . . . 109
7.4 Export graph data points to a file . . . 113
7.5 Copy/paste view output . . . 115

8 History log **116**

8.1 Element history 116

9 Batching and result handling **118**

9.1 How to handle results of analyses 118

III Bioinformatics **121**

10 Viewing and editing sequences **122**

10.1 View sequence 122

10.2 Circular DNA 130

10.3 Working with annotations 133

10.4 Element information 141

10.5 View as text 142

10.6 Creating a new sequence 142

10.7 Sequence Lists 143

11 Online database search **148**

11.1 GenBank search 148

11.2 UniProt (Swiss-Prot/TrEMBL) search 152

11.3 Search for structures at NCBI 154

11.4 Sequence web info 158

12 BLAST Search **160**

12.1 Running BLAST searches 161

12.2 Output from BLAST searches 167

12.3 Local BLAST databases 172

12.4 Manage BLAST databases 175

12.5 Bioinformatics explained: BLAST 176

13 3D molecule viewing **186**

13.1 Importing structure files 186

13.2 Viewing structure files 187

13.3 Selections and display of the 3D structure 188

13.4 3D Output 193

14 General sequence analyses **195**

14.1 Shuffle sequence 195
14.2 Dot plots 197
14.3 Local complexity plot 207
14.4 Sequence statistics 208
14.5 Join sequences 214
14.6 Pattern Discovery 215
14.7 Motif Search 217

15 Nucleotide analyses **225**

15.1 Convert DNA to RNA 225
15.2 Convert RNA to DNA 226
15.3 Reverse complements of sequences 227
15.4 Reverse sequence 228
15.5 Translation of DNA or RNA to protein 228
15.6 Find open reading frames 230

16 Protein analyses **233**

16.1 Signal peptide prediction 234
16.2 Protein charge 240
16.3 Transmembrane helix prediction 241
16.4 Antigenicity 242
16.5 Hydrophobicity 244
16.6 Pfam domain search 249
16.7 Secondary structure prediction 251
16.8 Protein report 253
16.9 Reverse translation from protein into DNA 255
16.10 Proteolytic cleavage detection 259

17 Restriction site analyses **265**

17.1 Dynamic restriction sites 266
17.2 Restriction site analysis from the Toolbox 270

17.3 Gel electrophoresis 275
17.4 Restriction enzyme lists 278

18 Sequence alignment 282
18.1 Create an alignment 283
18.2 View alignments 288
18.3 Edit alignments 292
18.4 Join alignments 294
18.5 Pairwise comparison 296
18.6 Bioinformatics explained: Multiple alignments 299

19 Phylogenetic trees 301
19.1 Inferring phylogenetic trees 301
19.2 Bioinformatics explained: phylogenetics 306

IV Appendix 310

A Comparison of workbenches 311

B Graph preferences 316

C Working with tables 318
C.1 Filtering tables 319

D BLAST databases 321
D.1 Peptide sequence databases 321
D.2 Nucleotide sequence databases 321
D.3 Adding more databases 322

E Proteolytic cleavage enzymes 324

F Formats for import and export 326
F.1 List of bioinformatic data formats 326
F.2 List of graphics data formats 329

G IUPAC codes for amino acids 330

H IUPAC codes for nucleotides **332**

I Custom codon frequency tables **333**

Bibliography **334**

V Index **339**

Part I

Introduction

Chapter 1

Introduction to *CLC Protein Workbench*

Contents

1.1 Contact information . . . 12
1.2 Download and installation . . . 12
1.2.1 Program download . . . 12
1.2.2 Installation on Microsoft Windows . . . 12
1.2.3 Installation on Mac OS X . . . 13
1.2.4 Installation on Linux with an installer . . . 14
1.2.5 Installation on Linux with an RPM-package . . . 15
1.3 System requirements . . . 15
1.4 Licenses . . . 15
1.4.1 Request an evaluation license . . . 16
1.4.2 Download a license . . . 19
1.4.3 Import a license from a file . . . 21
1.4.4 Upgrade license . . . 21
1.4.5 Configure license server connection . . . 24
1.4.6 Limited mode . . . 27
1.5 About CLC Workbenches . . . 27
1.5.1 New program feature request . . . 28
1.5.2 Report program errors . . . 28
1.5.3 CLC Sequence Viewer vs. Workbenches . . . 29
1.6 When the program is installed: Getting started . . . 29
1.6.1 Quick start . . . 29
1.6.2 Import of example data . . . 30
1.7 Plug-ins . . . 30
1.7.1 Installing plug-ins . . . 30
1.7.2 Uninstalling plug-ins . . . 31
1.7.3 Updating plug-ins . . . 32
1.7.4 Resources . . . 32
1.8 Network configuration . . . 33
1.9 The format of the user manual . . . 34
1.9.1 Text formats . . . 35

Welcome to *CLC Protein Workbench* — a software package supporting your daily bioinformatics work.

We strongly encourage you to read this user manual in order to get the best possible basis for working with the software package.

This software is for research purposes only.

1.1 Contact information

The *CLC Protein Workbench* is developed by:

CLC bio A/S
Science Park Aarhus
Finlandsgade 10-12
8200 Aarhus N
Denmark

http://www.clcbio.com

VAT no.: DK 28 30 50 87

Telephone: +45 70 22 55 09
Fax: +45 70 22 55 19

E-mail: info@clcbio.com

If you have questions or comments regarding the program, you are welcome to contact our support function:

E-mail: support@clcbio.com

1.2 Download and installation

The *CLC Protein Workbench* is developed for Windows, Mac OS X and Linux. The software for either platform can be downloaded from http://www.clcbio.com/download.

1.2.1 Program download

The program is available for download on http://www.clcbio.com/download.

Before you download the program you are asked to fill in the **Download** dialog.

In the dialog you must choose:

- Which operating system you use
- Whether you would like to receive information about future releases

Depending on your operating system and your Internet browser, you are taken through some download options.

When the download of the installer (an application which facilitates the installation of the program) is complete, follow the platform specific instructions below to complete the installation procedure. [1]

1.2.2 Installation on Microsoft Windows

Starting the installation process is done in one of the following ways:

[1] You must be connected to the Internet throughout the installation process.

If you have downloaded an installer:
Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.

If you are installing from a CD:
Insert the CD into your CD-ROM drive.

Choose the "Install *CLC Protein Workbench*" from the menu displayed.

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.
- Choose where you would like to install the application and click **Next**.
- Choose a name for the Start Menu folder used to launch *CLC Protein Workbench* and click **Next**.
- Choose if *CLC Protein Workbench* should be used to open CLC files and click **Next**.
- Choose where you would like to create shortcuts for launching *CLC Protein Workbench* and click **Next**.
- Choose if you would like to associate .clc files to *CLC Protein Workbench*. If you check this option, double-clicking a file with a "clc" extension will open the *CLC Protein Workbench*.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Protein Workbench* right away, and click **Finish**.

When the installation is complete the program can be launched from the Start Menu or from one of the shortcuts you chose to create.

1.2.3 Installation on Mac OS X

Starting the installation process is done in one of the following ways:

If you have downloaded an installer:
Locate the downloaded installer and double-click the icon.
The default location for downloaded files is your desktop.

If you are installing from a CD:
Insert the CD into your CD-ROM drive and open it by double-clicking on the CD icon on your desktop.

Launch the installer by double-clicking on the "*CLC Protein Workbench*" icon.

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.

- Choose where you would like to install the application and click **Next**.
- Choose if *CLC Protein Workbench* should be used to open CLC files and click **Next**.
- Choose whether you would like to create desktop icon for launching *CLC Protein Workbench* and click **Next**.
- Choose if you would like to associate .clc files to *CLC Protein Workbench*. If you check this option, double-clicking a file with a "clc" extension will open the *CLC Protein Workbench*.
- Wait for the installation process to complete, choose whether you would like to launch *CLC Protein Workbench* right away, and click **Finish**.

When the installation is complete the program can be launched from your Applications folder, or from the desktop shortcut you chose to create. If you like, you can drag the application icon to the dock for easy access.

1.2.4 Installation on Linux with an installer

Navigate to the directory containing the installer and execute it. This can be done by running a command similar to:

```
# sh CLCProteinWorkbench_5_JRE.sh
```

If you are installing from a CD the installers are located in the "linux" directory.

Installing the program is done in the following steps:

- On the welcome screen, click **Next**.
- Read and accept the License agreement and click **Next**.
- Choose where you would like to install the application and click **Next**.
  For a system-wide installation you can choose for example /opt or /usr/local. If you do not have root privileges you can choose to install in your home directory.
- Choose where you would like to create symbolic links to the program
  DO NOT create symbolic links in the same location as the application.
  Symbolic links should be installed in a location which is included in your environment PATH. For a system-wide installation you can choose for example /usr/local/bin. If you do not have root privileges you can create a 'bin' directory in your home directory and install symbolic links there. You can also choose not to create symbolic links.
- Wait for the installation process to complete and click **Finish**.

If you choose to create symbolic links in a location which is included in your PATH, the program can be executed by running the command:

```
# clcproteinwb5
```

Otherwise you start the application by navigating to the location where you choose to install it and running the command:

```
# ./clcproteinwb5
```

1.2.5 Installation on Linux with an RPM-package

Navigate to the directory containing the rpm-package and install it using the rpm-tool by running a command similar to:

```
# rpm -ivh CLCProteinWorkbench_5_JRE.rpm
```

If you are installing from a CD the rpm-packages are located in the "RPMS" directory.
Installation of RPM-packages usually requires root-privileges.

When the installation process is finished the program can be executed by running the command:

```
# clcproteinwb5
```

1.3 System requirements

The system requirements of *CLC Protein Workbench* are these:

- Windows XP, Windows Vista, or Windows 7, Windows Server 2003 or Windows Server 2008
- Mac OS X 10.5 or newer. PowerPC G4, G5 or Intel CPU required.
- Linux: RedHat 5 or later. SuSE 10 or later.
- 32 or 64 bit
- 256 MB RAM required
- 512 MB RAM recommended
- 1024 x 768 display recommended

1.4 Licenses

When you have installed *CLC Protein Workbench*, and start for the first time, you will meet the license assistant, shown in figure 1.1.

The following options are available. They will be described in detail in the following sections.

- **Request an evaluation license**. The license is a fully functional, time-limited license (see below).
- **Download a license**. When you purchase a license, you will get a license ID from CLC bio. Using this option, you will get a license based on this ID.
- **Import a license from a file**. If CLC bio has provided a license file, or if you have downloaded a license from our web-based licensing system, you can import it using this option.
- **Upgrade license**. If you already have used a previous version of *CLC Protein Workbench*, and you are entitled to upgrading to the new *CLC Protein Workbench 5.8*, select this option to get a license upgrade.

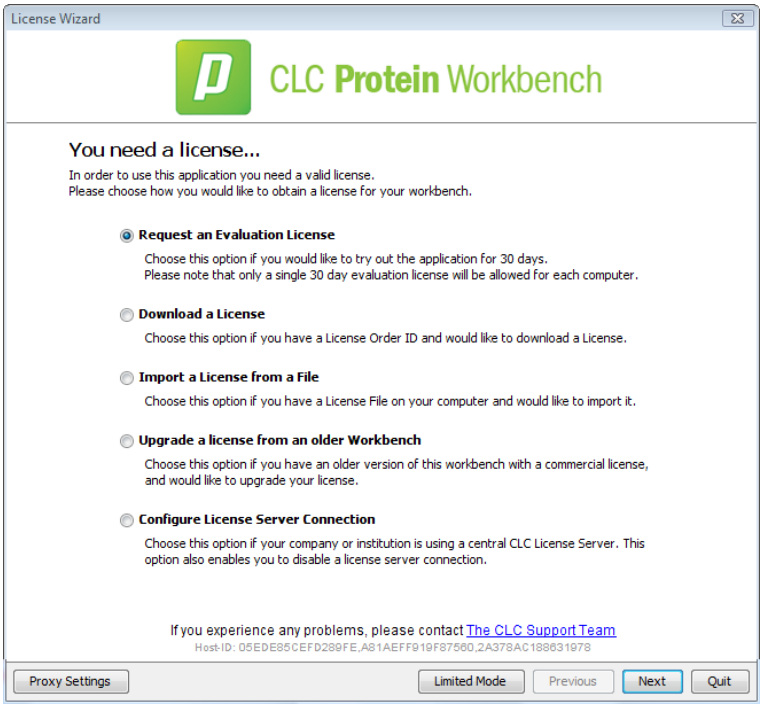

License Wizard
CLC Protein Workbench
You need a license...
In order to use this application you need a valid license.
Please choose how you would like to obtain a license for your workbench.
Request an Evaluation License
Choose this option if you would like to try out the application for 30 days.
Please note that only a single 30 day evaluation license will be allowed for each computer.
Download a License
Choose this option if you have a License Order ID and would like to download a License.
Import a License from a File
Choose this option if you have a License File on your computer and would like to import it.
Upgrade a license from an older Workbench
Choose this option if you have an older version of this workbench with a commercial license, and would like to upgrade your license.
Configure License Server Connection
Choose this option if your company or institution is using a central CLC License Server. This option also enables you to disable a license server connection.
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Limited Mode
Previous
Next
Quit

Figure 1.1: *The license assistant showing you the options for getting started.*

- **Configure license server connection**. If your organization has a license server, select this option to connect to the server.

Select an appropriate option and click **Next**.

If for some reason you don't have access to getting a license, you can click the **Limited Mode** button (see section 1.4.6).

1.4.1 Request an evaluation license

We offer a fully functional demo version of *CLC Protein Workbench* to all users, free of charge.

Each user is entitled to 30 days demo of *CLC Protein Workbench*. If you need more time for evaluating, another two weeks of demo can be requested.

We use the concept of *"quid quo pro"*. The last two weeks of free demo time given to you is therefore accompanied by a short-form questionnaire where you have the opportunity to give us feedback about the program.

The 30 days demo is offered for each major release of *CLC Protein Workbench*. You will therefore have the opportunity to try the next major version when it is released. (If you purchase *CLC Protein Workbench* the first year of updates is included.)

When you select to request an evaluation license, you will see the dialog shown in figure 1.2.

In this dialog, there are two options:

- **Direct download**. The workbench will attempt to contact the online CLC Licenses Service, and download the license directly. This method requires internet access from the workbench.

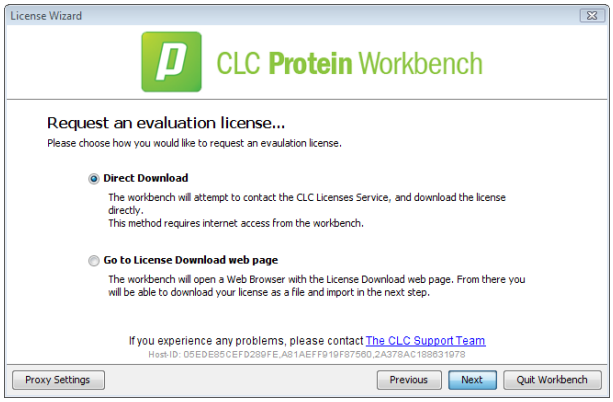

License Wizard
CLC Protein Workbench
Request an evaluation license...
Please choose how you would like to request an evaulation license.
Direct Download
The workbench will attempt to contact the CLC Licenses Service, and download the license directly.
This method requires internet access from the workbench.
Go to License Download web page
The workbench will open a Web Browser with the License Download web page. From there you will be able to download your license as a file and import in the next step.
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.2: *Choosing between direct download or download web page.*

- **Go to license download web page**. The workbench will open a Web Browser with the License Download web page when you click **Next**. From there you will be able to download your license as a file and import it. This option allows you to get a license, even though the Workbench does not have direct access to the CLC Licenses Service.

If you select the first option, and it turns out that you do not have internet access from the Workbench (because of a firewall, proxy server etc.), you will be able to click **Previous** and use the other option instead.

Direct download

Selecting the first option takes you to the dialog shown in figure 1.3.

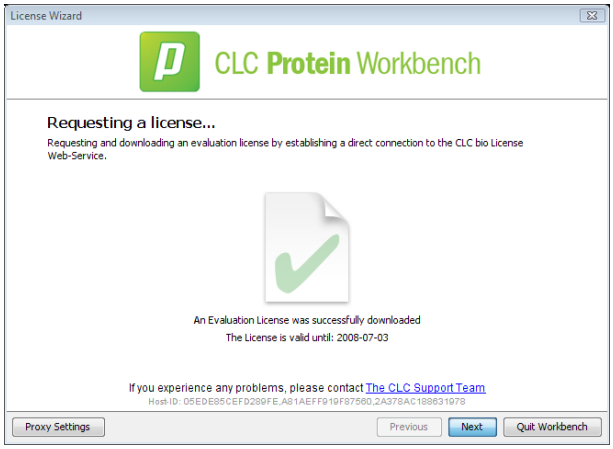

License Wizard
CLC Protein Workbench
Requesting a license...
Requesting and downloading an evaluation license by establishing a direct connection to the CLC bio License Web-Service.
An Evaluation License was successfully downloaded
The License is valid until: 2008-07-03
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.3: *A license has been downloaded.*

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

Go to license download web page

Selecting the second option, *Go to license download web page*, opens the license web page as shown in 1.4.

Click the **Request Evaluation License** button, and you will be able to save the license on your computer, e.g. on the Desktop.

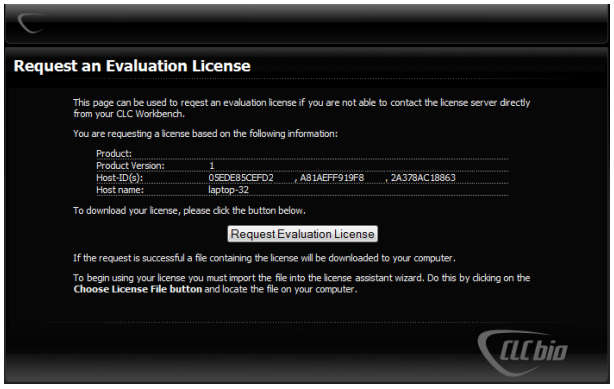

Request an Evaluation License
This page can be used to reqest an evaluation license if you are not able to contact the license server directly from your CLC Workbench.
You are requesting a license based on the following information:
Product:
Product Version: 1
Host-ID(s): 05EDE85CEFD2 , A81AEFF919F8 , 2A378AC18863
Host name: laptop-32
To download your license, please click the button below.
Request Evaluation License
If the request is successful a file containing the license will be downloaded to your computer.
To begin using your license you must import the file into the license assistant wizard. Do this by clicking on the Choose License File button and locate the file on your computer.
CLC bio

Figure 1.4: *The license web page where you can download a license.*

Back in the Workbench window, you will now see the dialog shown in 1.5.

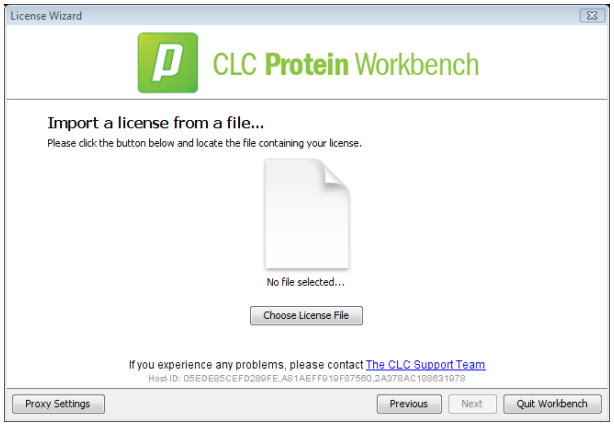

License Wizard
CLC Protein Workbench
Import a license from a file...
Please click the button below and locate the file containing your license.
No file selected...
Choose License File
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.5: *Importing the license downloaded from the web site.*

Click the **Choose License File** button and browse to find the license file you saved before (e.g. on your Desktop). When you have selected the file, click **Next**.

Accepting the license agreement

Regardless of which option you chose above, you will now see the dialog shown in figure 1.6.

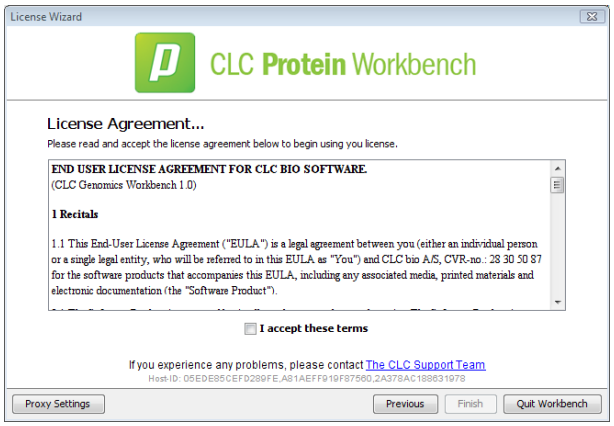

License Wizard
CLC Protein Workbench
License Agreement...
Please read and accept the license agreement below to begin using you license.
END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE
(CLC Genomics Workbench 1.0)
1 Recitals
1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").
I accept these terms
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Finish
Quit Workbench

Figure 1.6: *Read the license agreement carefully.*

Please read the License agreement carefully before clicking **I accept these terms** and **Finish**.

1.4.2 Download a license

When you purchase a license, you will get a license ID from CLC bio. Using this option, you will get a license based on this ID. When you have clicked **Next**, you will see the dialog shown in 1.7. At the top, enter the ID (paste using Ctrl+V or ⌘ + V on Mac).

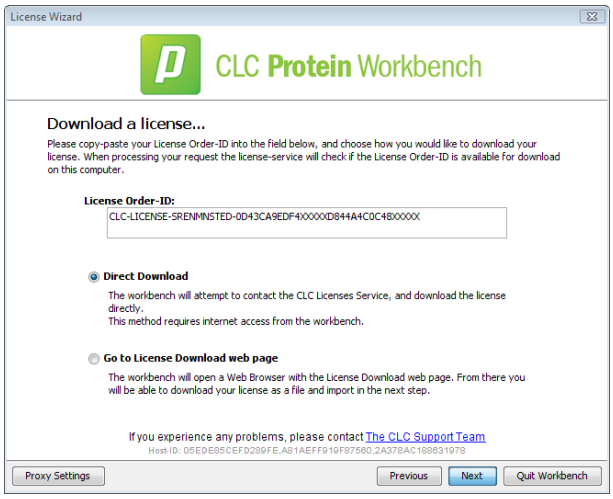

License Wizard
CLC Protein Workbench
Download a license...
Please copy-paste your License Order-ID into the field below, and choose how you would like to download your license. When processing your request the license-service will check if the License Order-ID is available for download on this computer.
License Order-ID:
CLC-LICENSE-SRENMNSTED-0D43CA9EDF4XXXXXD844A4C0C48XXXXX
Direct Download
The workbench will attempt to contact the CLC Licenses Service, and download the license directly.
This method requires internet access from the workbench.
Go to License Download web page
The workbench will open a Web Browser with the License Download web page. From there you will be able to download your license as a file and import in the next step.
If you experience any problems, please contact The CLC Support Team
Host ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.7: *Entering a license ID provided by CLC bio (the license ID in this example is artificial).*

In this dialog, there are two options:

- **Direct download**. The workbench will attempt to contact the online CLC Licenses Service, and download the license directly. This method requires internet access from the workbench.
- **Go to license download web page**. The workbench will open a Web Browser with the License Download web page when you click **Next**. From there you will be able to download your license as a file and import it. This option allows you to get a license, even though the Workbench does not have direct access to the CLC Licenses Service.

If you select the first option, and it turns out that you do not have internet access from the Workbench (because of a firewall, proxy server etc.), you will be able to click **Previous** and use the other option instead.

Direct download

Selecting the first option takes you to the dialog shown in figure 1.8.

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

Go to license download web page

Selecting the second option, *Go to license download web page*, opens the license web page as shown in 1.9.

Click the **Request Evaluation License** button, and you will be able to save the license on your computer, e.g. on the Desktop.

Back in the Workbench window, you will now see the dialog shown in 1.10.

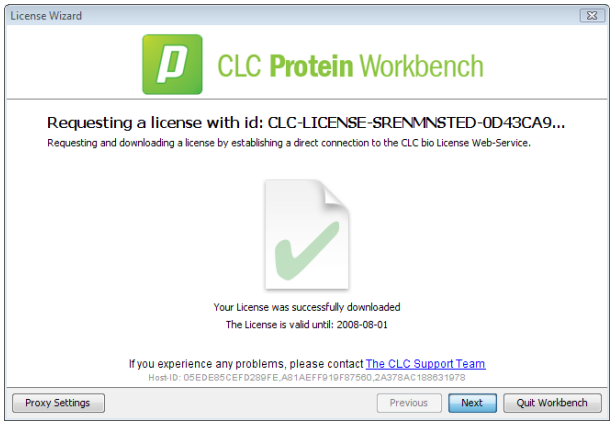

License Wizard
CLC Protein Workbench
Requesting a license with id: CLC-LICENSE-SRENMNSTED-0D43CA9...
Requesting and downloading a license by establishing a direct connection to the CLC bio License Web-Service.
Your License was successfully downloaded
The License is valid until: 2008-08-01
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.8: *A license has been downloaded.*

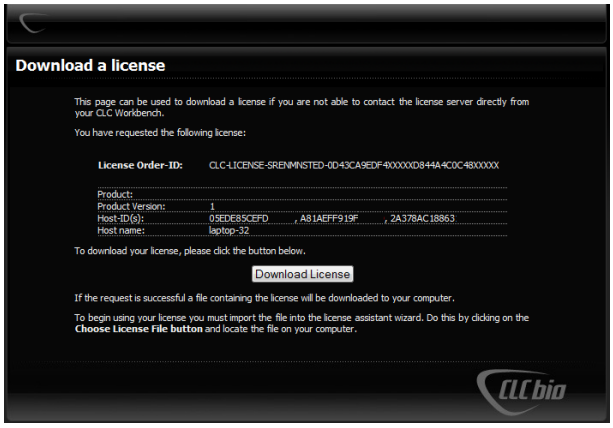

Download a license
This page can be used to download a license if you are not able to contact the license server directly from your CLC Workbench.
You have requested the following license:
License Order-ID: CLC-LICENSE-SRENMNSTED-0D43CA9EDF4XXXXD844A4C0C48XXXXX
Product:
Product Version: 1
Host-ID(s): 05EDE85CEFD , A81AEFF919F , 2A378AC18863
Host name: laptop-32
To download your license, please click the button below.
Download License
If the request is successful a file containing the license will be downloaded to your computer.
To begin using your license you must import the file into the license assistant wizard. Do this by clicking on the Choose License File button and locate the file on your computer.
CLC bio

Figure 1.9: *The license web page where you can download a license.*

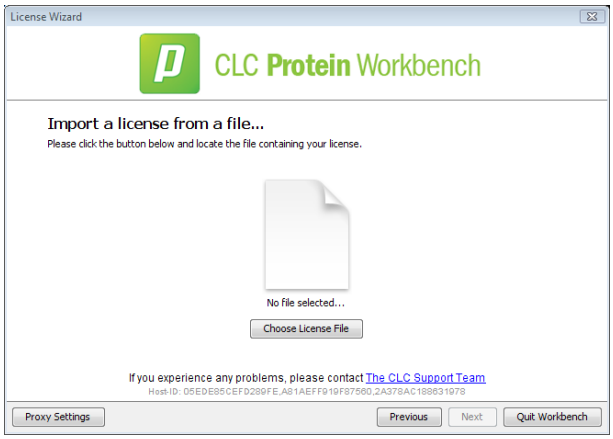

License Wizard
CLC Protein Workbench
Import a license from a file...
Please click the button below and locate the file containing your license.
No file selected...
Choose License File
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.10: *Importing the license downloaded from the web site.*

Click the **Choose License File** button and browse to find the license file you saved before (e.g. on your Desktop). When you have selected the file, click **Next**.

Accepting the license agreement

Regardless of which option you chose above, you will now see the dialog shown in figure 1.11.

Please read the License agreement carefully before clicking **I accept these terms** and **Finish**.

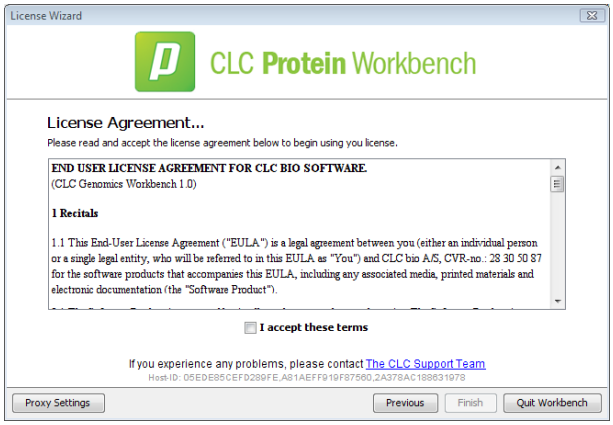

License Wizard
CLC Protein Workbench
License Agreement...
Please read and accept the license agreement below to begin using you license.
END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)
1 Recitals
1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").
I accept these terms
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Finish
Quit Workbench

Figure 1.11: *Read the license agreement carefully.*

1.4.3 Import a license from a file

If you are provided a license file instead of a license ID, you will be able to import the file using this option.

When you have clicked **Next**, you will see the dialog shown in 1.12.

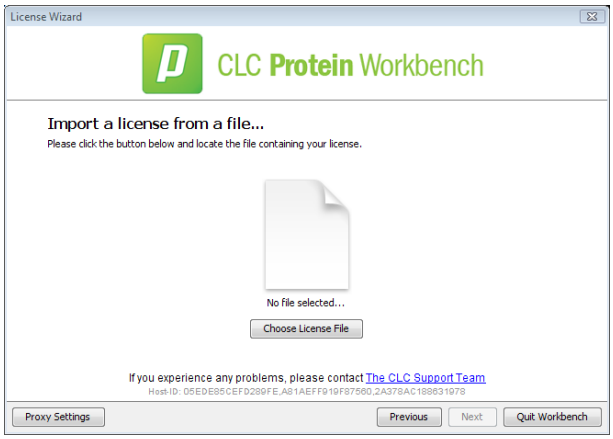

License Wizard
CLC Protein Workbench
Import a license from a file...
Please click the button below and locate the file containing your license.
No file selected...
Choose License File
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.12: *Selecting a license file .*

Click the **Choose License File** button and browse to find the license file provided by CLC bio. When you have selected the file, click **Next**.

Accepting the license agreement

Regardless of which option you chose above, you will now see the dialog shown in figure 1.13.

Please read the License agreement carefully before clicking **I accept these terms** and **Finish**.

1.4.4 Upgrade license

If you already have used a previous version of *CLC Protein Workbench*, and you are entitled to upgrading to the new *CLC Protein Workbench 5.8*, select this option to get a license upgrade.

When you click **Next**, the workbench will search for a previous installation of *CLC Protein Workbench*. It will then locate the old license.

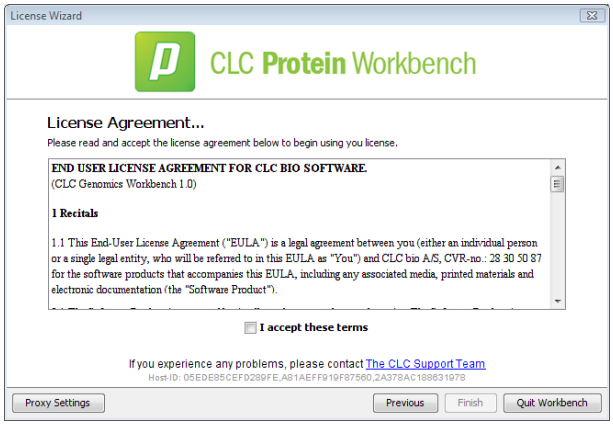

License Wizard
CLC Protein Workbench
License Agreement...
Please read and accept the license agreement below to begin using you license.
END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE
(CLC Genomics Workbench 1.0)
1 Recitals
1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").
I accept these terms
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Finish
Quit Workbench

Figure 1.13: *Read the license agreement carefully.*

If the Workbench succeeds to find an existing license, the next dialog will look as shown in figure 1.14.

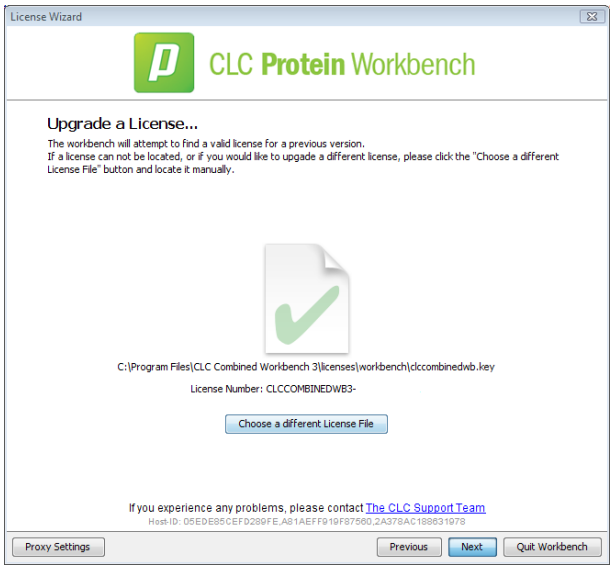

License Wizard
CLC Protein Workbench
Upgrade a License...
The workbench will attempt to find a valid license for a previous version.
If a license can not be located, or if you would like to upgade a different license, please click the "Choose a different License File" button and locate it manually.
C:\Program Files\CLC Combined Workbench 3\licenses\workbench\clccombinedwb.key
License Number: CLCCOMBINEDWB3-
Choose a different License File
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.14: *An old license is detected.*

When you click **Next**, the Workbench checks on CLC bio's web server to see if you are entitled to upgrade your license.

Note! If you should be entitled to get an upgrade, and you do not get one automatically in this process, please contact support@clcbio.com.

In this dialog, there are two options:

- **Direct download**. The workbench will attempt to contact the online CLC Licenses Service, and download the license directly. This method requires internet access from the workbench.
- **Go to license download web page**. The workbench will open a Web Browser with the License Download web page when you click **Next**. From there you will be able to download your license as a file and import it. This option allows you to get a license, even though the Workbench does not have direct access to the CLC Licenses Service.

If you select the first option, and it turns out that you do not have internet access from the

Workbench (because of a firewall, proxy server etc.), you will be able to click **Previous** and use the other option instead.

Direct download

Selecting the first option takes you to the dialog shown in figure 1.15.

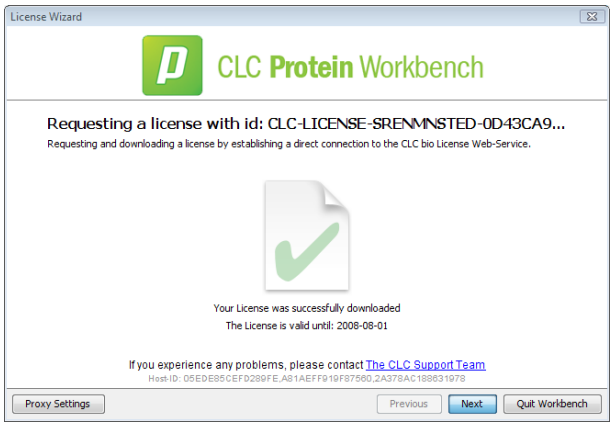

License Wizard
CLC Protein Workbench
Requesting a license with id: CLC-LICENSE-SRENMNSTED-0D43CA9...
Requesting and downloading a license by establishing a direct connection to the CLC bio License Web-Service.
Your License was successfully downloaded
The License is valid until: 2008-08-01
If you experience any problems, please contact The CLC Support Team
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.15: *A license has been downloaded.*

A progress for getting the license is shown, and when the license is downloaded, you will be able to click **Next**.

Go to license download web page

Selecting the second option, *Go to license download web page*, opens the license web page as shown in 1.16.

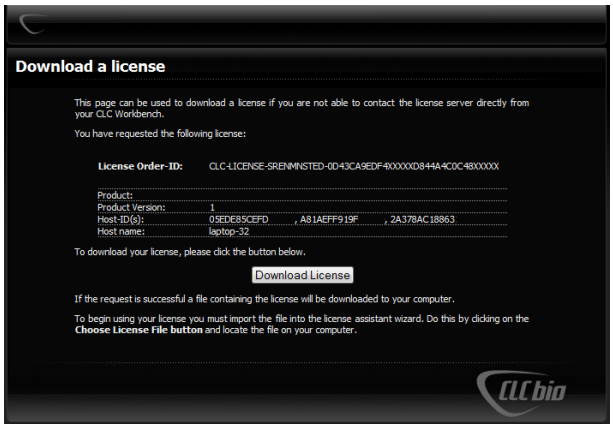

Download a license
This page can be used to download a license if you are not able to contact the license server directly from your CLC Workbench.
You have requested the following license:
License Order-ID:
Product:
Product Version:
Host-ID(s):
Host name:
laptop-32
To download your license, please click the button below.
Download License
If the request is successful a file containing the license will be downloaded to your computer.
To begin using your license you must import the file into the license assistant wizard. Do this by clicking on the Choose License File button and locate the file on your computer.

Figure 1.16: *The license web page where you can download a license.*

Click the **Request Evaluation License** button, and you will be able to save the license on your computer, e.g. on the Desktop.

Back in the Workbench window, you will now see the dialog shown in 1.17.

Click the **Choose License File** button and browse to find the license file you saved before (e.g. on your Desktop). When you have selected the file, click **Next**.

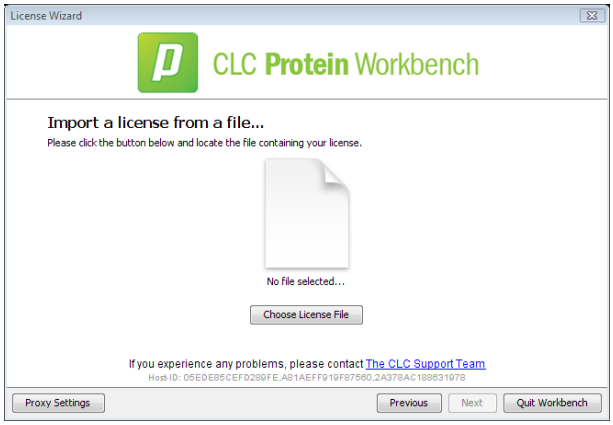

License Wizard
CLC Protein Workbench
Import a license from a file...
Please click the button below and locate the file containing your license.
No file selected...
Choose License File
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Next
Quit Workbench

Figure 1.17: *Importing the license downloaded from the web site.*

Accepting the license agreement

Regardless of which option you chose above, you will now see the dialog shown in figure 1.18.

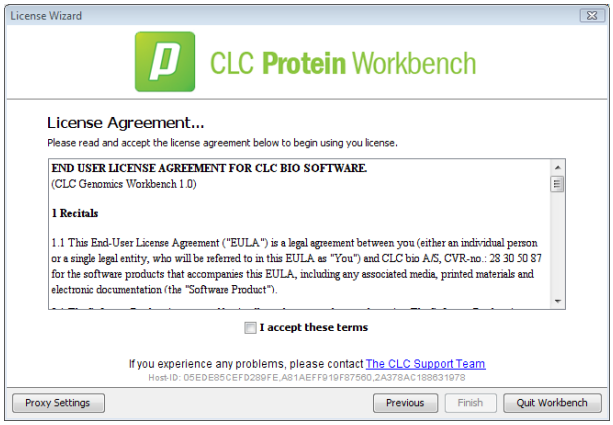

License Wizard
CLC Protein Workbench
License Agreement...
Please read and accept the license agreement below to begin using you license.
END USER LICENSE AGREEMENT FOR CLC BIO SOFTWARE.
(CLC Genomics Workbench 1.0)
1 Recitals
1.1 This End-User License Agreement ("EULA") is a legal agreement between you (either an individual person or a single legal entity, who will be referred to in this EULA as "You") and CLC bio A/S, CVR-no.: 28 30 50 87 for the software products that accompanies this EULA, including any associated media, printed materials and electronic documentation (the "Software Product").
I accept these terms
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
Proxy Settings
Previous
Finish
Quit Workbench

Figure 1.18: *Read the license agreement carefully.*

Please read the License agreement carefully before clicking **I accept these terms** and **Finish**.

1.4.5 Configure license server connection

If your organization has installed a license server, you can use a floating license. The license server has a set of licenses that can be used on all computers on the network. If the server has e.g. 10 licenses, it means that maximum 10 computers can use a license *simultaneously*. When you have selected this option and click **Next**, you will see the dialog shown in figure 1.19.

This dialog lets you specify how to connect to the license server:

- **Connect to a license server**. Check this option if you wish to use the license server.
- **Automatically detect license server**. By checking this option you do not have to enter more information to connect to the server.
- **Manually specify license server**. There can be technical limitations which mean that the license server cannot be detected automatically, and in this case you need to specify more options manually:

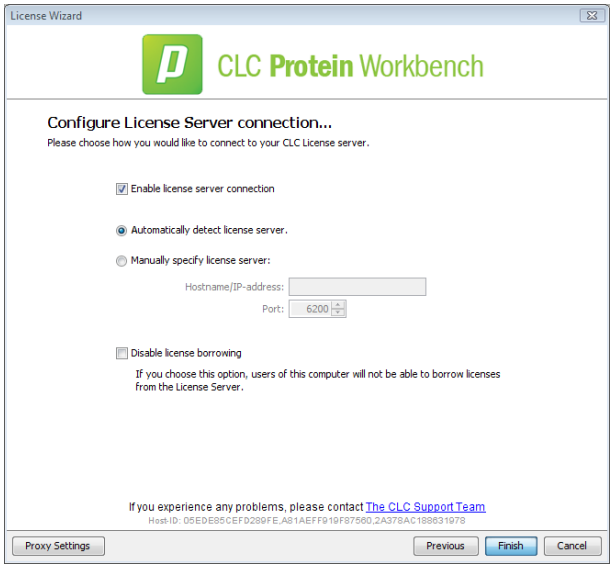

License Wizard
CLC Protein Workbench
Configure License Server connection...
Please choose how you would like to connect to your CLC License server.
Enable license server connection
Automatically detect license server.
Manually specify license server:
Hostname/IP-address:
Port:
6200
Disable license borrowing
If you choose this option, users of this computer will not be able to borrow licenses from the License Server.
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188831978
Proxy Settings
Previous
Finish
Cancel

Figure 1.19: *Connecting to a license server.*

  - **Host name**. Enter the address for the licenser server.
  - **Port**. Specify which port to use.

- **Disable license borrowing on this computer**. If you do not want users of the computer to borrow a license (see section 1.4.5), you can check this option.

Borrow a license

A floating license can only be used when you are connected to the license server. If you wish to use the *CLC Protein Workbench* when you are not connected to the server, you can *borrow* a license. Borrowing a license means that you take one of the floating licenses available on the server and borrow it for a specified amount of time. During this time period, there will be one less floating license available on the server.

At the point where you wish to borrow a license, you have to be connected to the license server. The procedure for borrowing is this:

1. Click **Help** | **License Manager** to display the dialog shown in figure 1.22.
2. Use the checkboxes to select the license(s) that you wish to borrow.
3. Select how long time you wish to borrow the license, and click **Borrow Licenses**.
4. You can now go offline and work with *CLC Protein Workbench*.
5. When the borrow time period has elapsed, you have to connect to the license server again to use *CLC Protein Workbench*.
6. When the borrow time period has elapsed, the license server will make the floating license available for other users.

Note that the time period is not the period of time that you actually use the Workbench.

Note! When your organization's license server is installed, license borrowing can be turned off. In that case, you will not be able to borrow licenses.

No license available...

If all the licenses on the server are in use, you will see a dialog as shown in figure 1.20 when you start the Workbench.

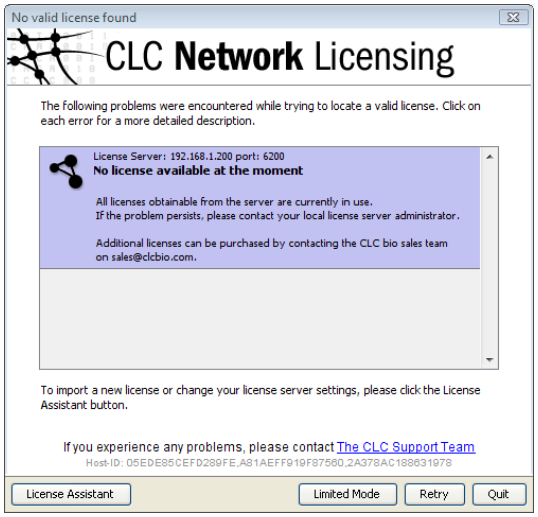

No valid license found
CLC Network Licensing
The following problems were encountered while trying to locate a valid license. Click on each error for a more detailed description.
License Server: 192.168.1.200 port: 6200
No license available at the moment
All licenses obtainable from the server are currently in use.
If the problem persists, please contact your local license server administrator.
Additional licenses can be purchased by contacting the CLC bio sales team on sales@clcbio.com.
To import a new license or change your license server settings, please click the License Assistant button.
If you experience any problems, please contact The CLC Support Team
Host-ID: 05EDE85CEFD289FE,A81AEFF919F87560,2A378AC188631978
License Assistant
Limited Mode
Retry
Quit

Figure 1.20: *No more licenses available on the server.*

In this case, please contact your organization's license server administrator. To purchase additional licenses, contact sales@clcbio.com.

You can also click the **Limited Mode** button (see section 1.4.6).

If your connection to the license server is lost, you will see a dialog as shown in figure 1.21.

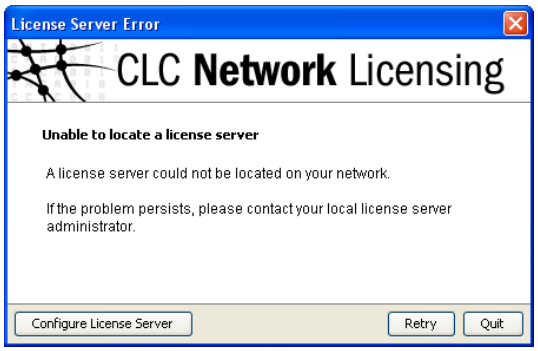

License Server Error
CLC Network Licensing
Unable to locate a license server
A license server could not be located on your network.
If the problem persists, please contact your local license server administrator.
Configure License Server
Retry
Quit

Figure 1.21: *Unable to contact license server.*

In this case, you need to make sure that you have access to the license server, and that the server is running. However, there may be situations where you wish to use another license, or see information about the license you currently use. In this case, open the license manager:

Help | License Manager ( )

The license manager is shown in figure 1.22.

Besides letting you borrow licenses (see section 1.4.5), this dialog can be used to:

- See information about the license (e.g. what kind of license, when it expires)

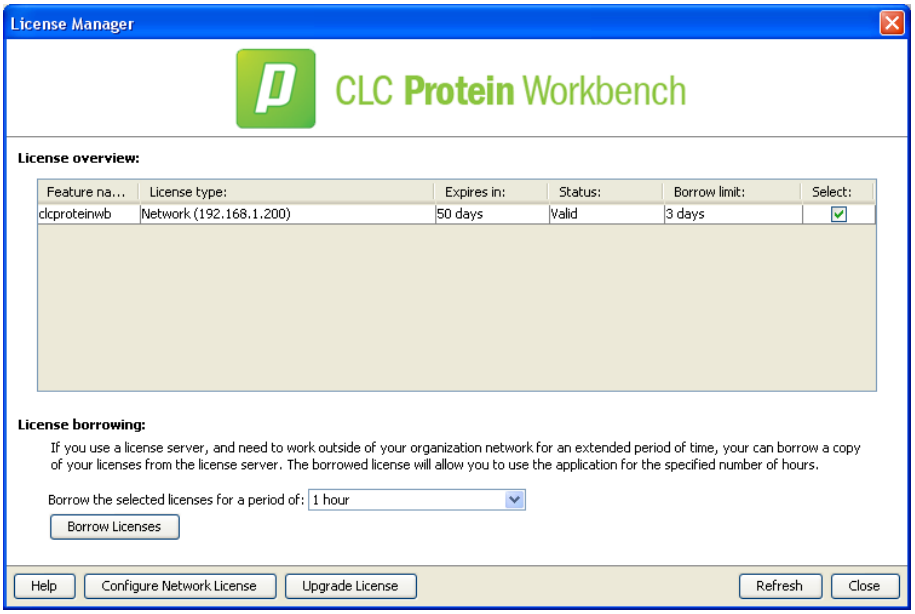

License Manager
CLC Protein Workbench
License overview:
Feature na...
License type:
Expires in:
Status:
Borrow limit:
Select:
clcproteinwb
Network (192.168.1.200)
50 days
Valid
3 days
License borrowing:
If you use a license server, and need to work outside of your organization network for an extended period of time, your can borrow a copy of your licenses from the license server. The borrowed license will allow you to use the application for the specified number of hours.
Borrow the selected licenses for a period of:
1 hour
Borrow Licenses
Help
Configure Network License
Upgrade License
Refresh
Close

Figure 1.22: *The license manager.*

- Configure how to connect to a license server (**Configure License Server** the button at the lower left corner). Clicking this button will display a dialog similar to figure 1.19.
- Upgrade from an evaluation license by clicking the **Upgrade license** button. This will display the dialog shown in figure 1.1.

If you wish to switch away from using a floating license, click **Configure License Server** and choose not to connect to a license server in the dialog. When you restart *CLC Protein Workbench*, you will be asked for a license as described in section 1.4.

1.4.6 Limited mode

We have created the limited mode to prevent a situation where you are unable to access your data because you do not have a license. When you run in limited mode, a lot of the tools in the Workbench are not available, but you still have access to your data (also when stored in a *CLC Bioinformatics Database*). When running in limited mode, the functionality is equivalent to the *CLC Sequence Viewer* (see section A).

To get out of the limited mode and run the Workbench normally, restart the Workbench. When you restart the Workbench will try to find a proper license and if it does, it will start up normally. If it can't find a license, you will again have the option of running in limited mode.

1.5 About CLC Workbenches

In November 2005 CLC bio released two Workbenches: *CLC Free Workbench* and *CLC Protein Workbench*. *CLC Protein Workbench* is developed from the free version, giving it the well-tested user friendliness and look & feel. However, the *CLC Protein Workbench* includes a range of more advanced analyses.

In March 2006, *CLC DNA Workbench* (formerly *CLC Gene Workbench*) and *CLC Main Workbench* were added to the product portfolio of CLC bio. Like *CLC Protein Workbench*, *CLC DNA Workbench*

builds on *CLC Free Workbench*. It shares some of the advanced product features of *CLC Protein Workbench*, and it has additional advanced features. *CLC Main Workbench* holds all basic and advanced features of the *CLC Workbenches*.

In June 2007, *CLC RNA Workbench* was released as a sister product of *CLC Protein Workbench* and *CLC DNA Workbench*. *CLC Main Workbench* now also includes all the features of *CLC RNA Workbench*.

In March 2008, the *CLC Free Workbench* changed name to *CLC Sequence Viewer*.

In June 2008, the first version of the *CLC Genomics Workbench* was released due to an extraordinary demand for software capable of handling sequencing data from the new high-throughput sequencing systems like 454, Illumina Genome Analyzer and SOLiD.

For an overview of which features all the applications include, see http://www.clcbio.com/features.

In December 2006, CLC bio released a **Software Developer Kit** which makes it possible for anybody with a knowledge of programming in Java to develop plug-ins. The plug-ins are fully integrated with the CLC Workbenches and the Viewer and provide an easy way to customize and extend their functionalities.

All our software will be improved continuously. If you are interested in receiving news about updates, you should register your e-mail and contact data on http://www.clcbio.com, if you haven't already registered when you downloaded the program.

1.5.1 New program feature request

The CLC team is continuously improving the *CLC Protein Workbench* with our users' interests in mind. Therefore, we welcome all requests and feedback from users, and hope suggest new features or more general improvements to the program on support@clcbio.com.

1.5.2 Report program errors

CLC bio is doing everything possible to eliminate program errors. Nevertheless, some errors might have escaped our attention. If you discover an error in the program, you can use the **Report a Program Error** function in the **Help** menu of the program to report it. In the **Report a Program Error** dialog you are asked to write your e-mail address (optional). This is because we would like to be able to contact you for further information about the error or for helping you with the problem.

Note! No personal information is sent via the error report. Only the information which can be seen in the **Program Error Submission Dialog** is submitted.

You can also write an e-mail to support@clcbio.com. Remember to specify how the program error can be reproduced.

All errors will be treated seriously and with gratitude.

We appreciate your help.

Start in safe mode

If the program becomes unstable on start-up, you can start it in **Safe mode**. This is done by pressing and holding down the Shift button while the program starts.

When starting in safe mode, the user settings (e.g. the settings in the **Side Panel**) are deleted and cannot be restored. Your data stored in the **Navigation Area** is not deleted. When started in safe mode, some of the functionalities are missing, and you will have to restart the *CLC Protein Workbench* again (without pressing Shift).

1.5.3 CLC Sequence Viewer vs. Workbenches

The advanced analyses of the commercial workbenches, *CLC Protein Workbench*, *CLC RNA Workbench* and *CLC DNA Workbench* are not present in *CLC Sequence Viewer*. Likewise, some advanced analyses are available in *CLC DNA Workbench* but not in *CLC RNA Workbench* or *CLC Protein Workbench*, and vice versa. All types of basic and advanced analyses are available in *CLC Main Workbench*.

However, the output of the commercial workbenches can be viewed in all other workbenches. This allows you to share the result of your advanced analyses from e.g. *CLC Main Workbench*, with people working with e.g. *CLC Sequence Viewer*. They will be able to view the results of your analyses, but not redo the analyses.

The CLC Workbenches and the CLC Sequence Viewer are developed for Windows, Mac and Linux platforms. Data can be exported/imported between the different platforms in the same easy way as when exporting/importing between two computers with e.g. Windows.

1.6 When the program is installed: Getting started

CLC Protein Workbench includes an extensive **Help** function, which can be found in the **Help** menu of the program's **Menu bar**. The **Help** can also be shown by pressing F1. The help topics are sorted in a table of contents and the topics can be searched.

We also recommend our **Online presentations** where a product specialist from CLC bio demonstrates our software. This is a very easy way to get started using the program. Read more about online presentations here: `http://clcbio.com/presentation`.

1.6.1 Quick start

When the program opens for the first time, the background of the workspace is visible. In the background are three quick start shortcuts, which will help you getting started. These can be seen in figure 1.23.

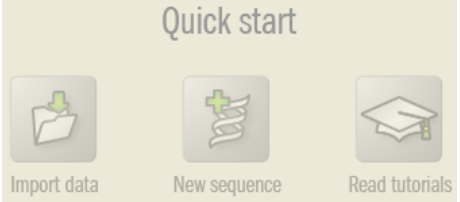


Figure 1.23: *Three available Quick start short cuts, available in the background of the workspace.*

The function of the three quick start shortcuts is explained here:

- **Import data.** Opens the **Import** dialog, which you let you browse for, and import data from your file system.
- **New sequence.** Opens a dialog which allows you to enter your own sequence.
- **Read tutorials.** Opens the tutorials menu with a number of tutorials. These are also available from the **Help** menu in the **Menu bar**.

1.6.2 Import of example data

It might be easier to understand the logic of the program by trying to do simple operations on existing data. Therefore *CLC Protein Workbench* includes an example data set.

When downloading *CLC Protein Workbench* you are asked if you would like to import the example data set. If you accept, the data is downloaded automatically and saved in the program. If you didn't download the data, or for some other reason need to download the data again, you have two options:

You can click **Install Example Data** ( ) in the **Help** menu of the program. This installs the data automatically. You can also go to `http://www.clcbio.com/download` and download the example data from there.

If you download the file from the website, you need to import it into the program. See chapter 7.1 for more about importing data.

1.7 Plug-ins

When you install *CLC Protein Workbench*, it has a standard set of features. However, you can upgrade and customize the program using a variety of plug-ins.

As the range of plug-ins is continuously updated and expanded, they will not be listed here. Instead we refer to `http://www.clcbio.com/plug-ins` for a full list of plug-ins with descriptions of their functionalities.

1.7.1 Installing plug-ins

Plug-ins are installed using the plug-in manager[2]:

Help in the Menu Bar | **Plug-ins and Resources... (** **)**

or **Plug-ins (** **) in the Toolbar**

The plug-in manager has four tabs at the top:

- **Manage Plug-ins.** This is an overview of plug-ins that are installed.
- **Download Plug-ins.** This is an overview of available plug-ins on CLC bio's server.

[2] In order to install plug-ins on Windows Vista, the Workbench must be run in administrator mode: Right-click the program shortcut and choose "Run as Administrator". Then follow the procedure described below.

- **Manage Resources.** This is an overview of resources that are installed.
- **Download Resources.** This is an overview of available resources on CLC bio's server.

To install a plug-in, click the **Download Plug-ins** tab. This will display an overview of the plug-ins that are available for download and installation (see figure 1.24).

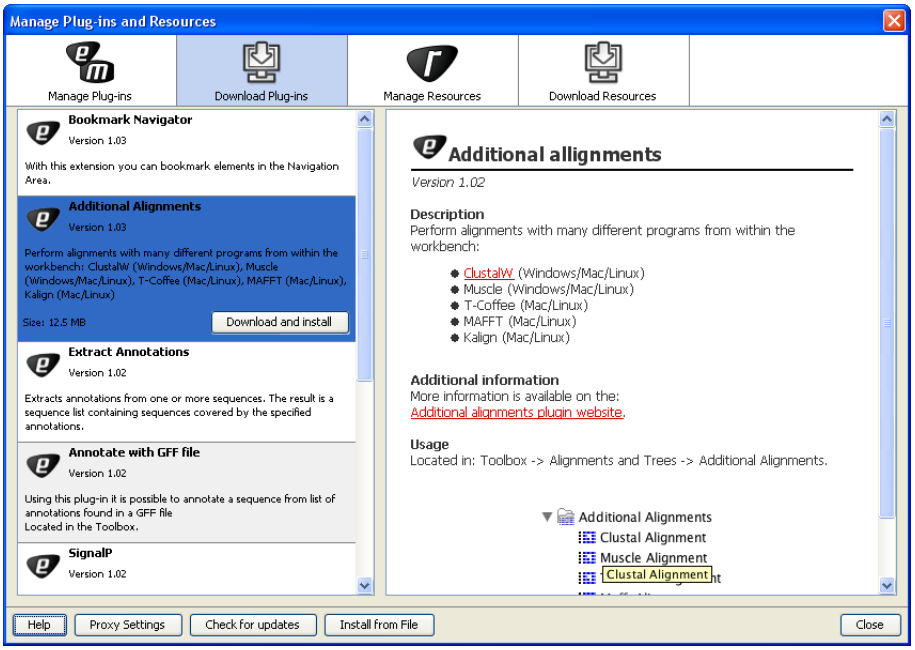

Manage Plug-ins and Resources
Manage Plug-ins
Download Plug-ins
Manage Resources
Download Resources
Bookmark Navigator
Version 1.03
With this extension you can bookmark elements in the Navigation Area.
Additional Alignments
Version 1.03
Perform alignments with many different programs from within the workbench: ClustalW (Windows/Mac/Linux), Muscle (Windows/Mac/Linux), T-Coffee (Mac/Linux), MAFFT (Mac/Linux), Kalign (Mac/Linux)
Size: 12.5 MB
Download and install
Extract Annotations
Version 1.02
Extracts annotations from one or more sequences. The result is a sequence list containing sequences covered by the specified annotations.
Annotate with GFF file
Version 1.02
Using this plug-in it is possible to annotate a sequence from list of annotations found in a GFF file
Located in the Toolbox.
SignalP
Version 1.02
Additional allignments
Version 1.02
Description
Perform alignments with many different programs from within the workbench:
ClustalW (Windows/Mac/Linux)
Muscle (Windows/Mac/Linux)
T-Coffee (Mac/Linux)
MAFFT (Mac/Linux)
Kalign (Mac/Linux)
Additional information
More information is available on the:
Additional alignments plugin website.
Usage
Located in: Toolbox -> Alignments and Trees -> Additional Alignments.
Additional Alignments
Clustal Alignment
Muscle Alignment
Clustal Alignment
Help
Proxy Settings
Check for updates
Install from File
Close

Figure 1.24: *The plug-ins that are available for download.*

Clicking a plug-in will display additional information at the right side of the dialog. This will also display a button: **Download and Install**.

Click the plug-in and press **Download and Install**. A dialog displaying progress is now shown, and the plug-in is downloaded and installed.

If the plug-in is not shown on the server, and you have it on your computer (e.g. if you have downloaded it from our web-site), you can install it by clicking the **Install from File** button at the bottom of the dialog. This will open a dialog where you can browse for the plug-in. The plug-in file should be a file of the type ".cpa".

When you close the dialog, you will be asked whether you wish to restart the *CLC Protein Workbench*. The plug-in will not be ready for use before you have restarted.

1.7.2 Uninstalling plug-ins

Plug-ins are uninstalled using the plug-in manager:

Help in the Menu Bar | **Plug-ins and Resources... (** **)**

or **Plug-ins (** **) in the Toolbar**

This will open the dialog shown in figure 1.25.

The installed plug-ins are shown in this dialog. To uninstall:

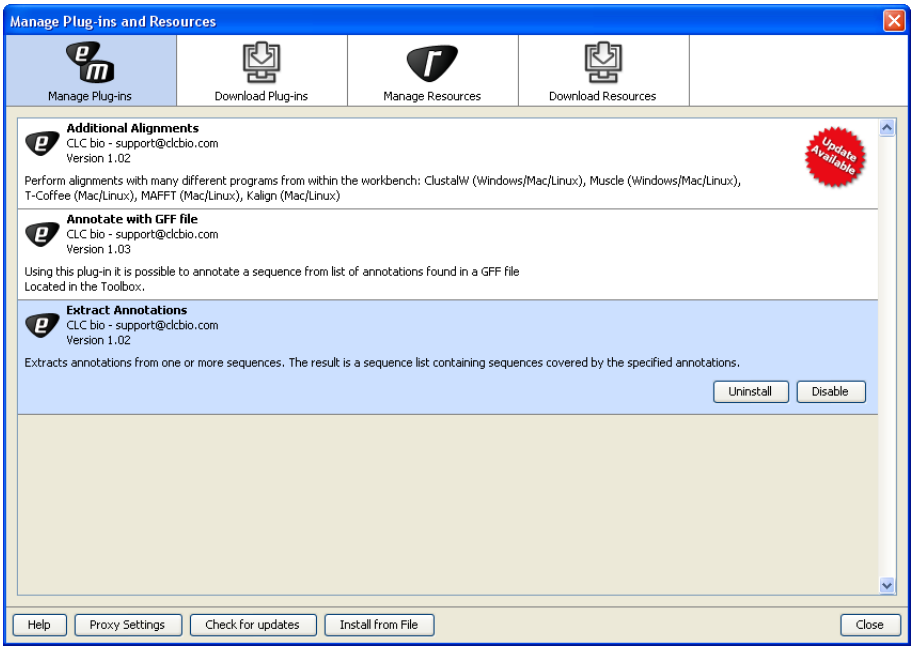

Manage Plug-ins and Resources
Manage Plug-ins
Download Plug-ins
Manage Resources
Download Resources
Additional Alignments
CLC bio - support@clcbio.com
Version 1.02
Update Available
Perform alignments with many different programs from within the workbench: ClustalW (Windows/Mac/Linux), Muscle (Windows/Mac/Linux), T-Coffee (Mac/Linux), MAFFT (Mac/Linux), Kalign (Mac/Linux)
Annotate with GFF file
CLC bio - support@clcbio.com
Version 1.03
Using this plug-in it is possible to annotate a sequence from list of annotations found in a GFF file
Located in the Toolbox.
Extract Annotations
CLC bio - support@clcbio.com
Version 1.02
Extracts annotations from one or more sequences. The result is a sequence list containing sequences covered by the specified annotations.
Uninstall
Disable
Help
Proxy Settings
Check for updates
Install from File
Close

Figure 1.25: *The plug-in manager with plug-ins installed.*

Click the plug-in | Uninstall

If you do not wish to completely uninstall the plug-in but you don't want it to be used next time you start the Workbench, click the **Disable** button.

When you close the dialog, you will be asked whether you wish to restart the workbench. The plug-in will not be uninstalled before the workbench is restarted.

1.7.3 Updating plug-ins

If a new version of a plug-in is available, you will get a notification during start-up as shown in figure 1.26.

In this list, select which plug-ins you wish to update, and click **Install Updates**. If you press **Cancel** you will be able to install the plug-ins later by clicking **Check for Updates** in the Plug-in manager (see figure 1.25).

1.7.4 Resources

Resources are downloaded, installed, un-installed and updated the same way as plug-ins. Click the **Download Resources** tab at the top of the plug-in manager, and you will see a list of available resources (see figure 1.27).

Currently, the only resources available are PFAM databases (for use with *CLC Protein Workbench* and *CLC Main Workbench*).

Because procedures for downloading, installation, uninstallation and updating are the same as for plug-ins see section 1.7.1 and section 1.7.2 for more information.

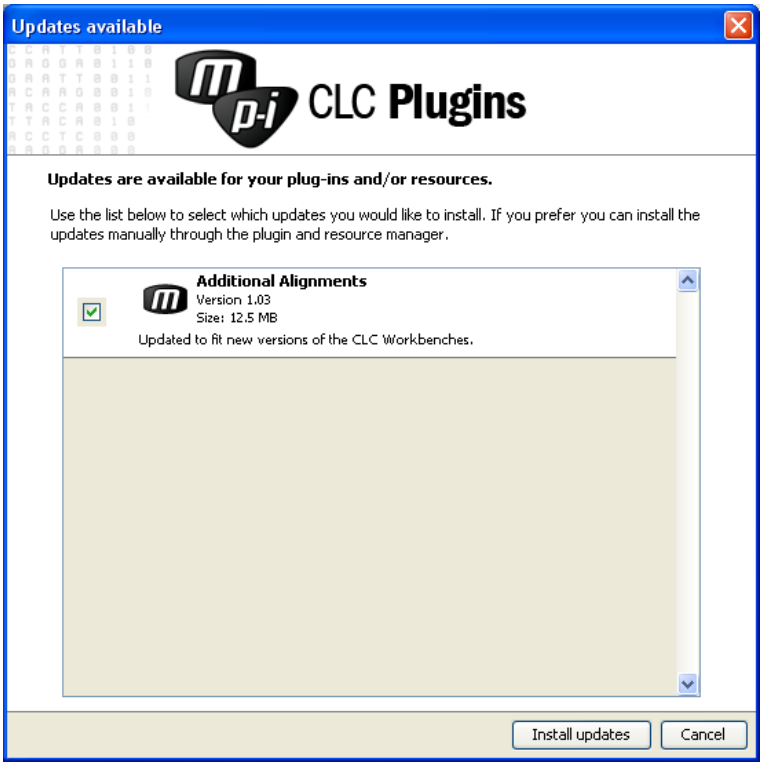

Updates available
CLC Plugins
Updates are available for your plug-ins and/or resources.
Use the list below to select which updates you would like to install. If you prefer you can install the updates manually through the plugin and resource manager.
Additional Alignments
Version 1.03
Size: 12.5 MB
Updated to fit new versions of the CLC Workbenches.
Install updates
Cancel

Figure 1.26: *Plug-in updates.*

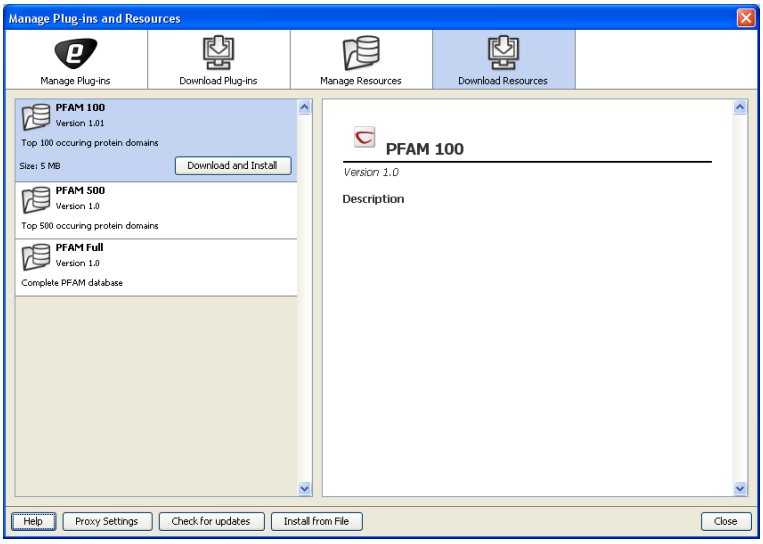


Figure 1.27: *Resources available for download.*

1.8 Network configuration

If you use a proxy server to access the Internet you must configure *CLC Protein Workbench* to use this. Otherwise you will not be able to perform any online activities (e.g. searching GenBank). *CLC Protein Workbench* supports the use of a HTTP-proxy and an anonymous SOCKS-proxy.

To configure your proxy settings, open *CLC Protein Workbench*, and go to the **Advanced**-tab of the **Preferences** dialog (figure 1.28) and enter the appropriate information. The **Preferences** dialog is opened from the **Edit** menu.

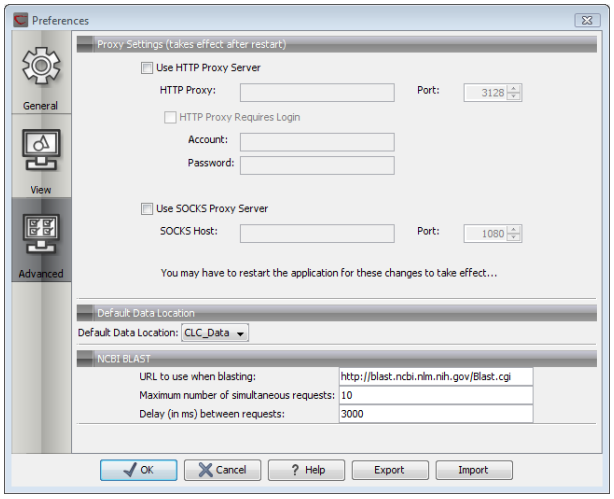


Figure 1.28: *Adjusting proxy preferences.*

You have the choice between a HTTP-proxy and a SOCKS-proxy. *CLC Protein Workbench* only supports the use of a SOCKS-proxy that does not require authorization.

Exclude hosts can be used if there are some hosts that should be contacted directly and not through the proxy server. The value can be a list of hosts, each separated by a `|`, and in addition a wildcard character `*` can be used for matching. For example: `*.foo.com|localhost`.

If you have any problems with these settings you should contact your systems administrator.

1.9 The format of the user manual

This user manual offers support to Windows, Mac OS X and Linux users. The software is very similar on these operating systems. In areas where differences exist, these will be described separately. However, the term "right-click" is used throughout the manual, but some Mac users may have to use Ctrl+click in order to perform a "right-click" (if they have a single-button mouse).

The most recent version of the user manuals can be downloaded from `http://www.clcbio.com/usermanuals`.

The user manual consists of four parts.

- The **first part** includes the introduction and some tutorials showing how to apply the most significant functionalities of *CLC Protein Workbench*.
- The **second part** describes in detail how to operate all the program's basic functionalities.
- The **third part** digs deeper into some of the bioinformatic features of the program. In this part, you will also find our "Bioinformatics explained" sections. These sections elaborate on the algorithms and analyses of *CLC Protein Workbench* and provide more general knowledge of bioinformatic concepts.
- The **fourth part** is the Appendix and Index.

Each chapter includes a short table of contents.

1.9.1 Text formats

In order to produce a clearly laid-out content in this manual, different formats are applied:

- A feature in the program is in bold starting with capital letters. (Example: **Navigation Area**)
- An explanation of how a particular function is activated, is illustrated by "|" and bold. (E.g.: **select the element** | **Edit** | **Rename**)

Chapter 2

Tutorials

Contents

2.1 Tutorial: Getting started . . . 37
2.1.1 Creating a a folder . . . 38
2.1.2 Import data . . . 38
2.2 Tutorial: View sequence . . . 38
2.3 Tutorial: Side Panel Settings . . . 40
2.3.1 Saving the settings in the Side Panel . . . 42
2.3.2 Applying saved settings . . . 43
2.4 Tutorial: GenBank search and download . . . 43
2.4.1 Searching for matching objects . . . 44
2.4.2 Saving the sequence . . . 44
2.5 Tutorial: BLAST search . . . 44
2.5.1 Performing the BLAST search . . . 45
2.5.2 Inspecting the results . . . 46
2.5.3 Using the BLAST table view . . . 47
2.6 Tutorial: Tips for specialized BLAST searches . . . 47
2.6.1 Locate a protein sequence on the chromosome . . . 48
2.6.2 BLAST for primer binding sites . . . 50
2.6.3 Finding remote protein homologues . . . 51
2.6.4 Further reading . . . 51
2.7 Tutorial: Proteolytic cleavage detection . . . 52
2.8 Tutorial: Align protein sequences . . . 54
2.8.1 The alignment dialog . . . 54
2.9 Tutorial: Create and modify a phylogenetic tree . . . 56
2.9.1 Tree layout . . . 56
2.10 Tutorial: Find restriction sites . . . 57
2.10.1 The Side Panel way of finding restriction sites . . . 57
2.10.2 The Toolbox way of finding restriction sites . . . 58

This chapter contains tutorials representing some of the features of *CLC Protein Workbench*. The first tutorials are meant as a short introduction to operating the program. The last tutorials give examples of how to use some of the main features of *CLC Protein Workbench*. **Watch video tutorials at `http://www.clcbio.com/tutorials`.**

2.1 Tutorial: Getting started

This brief tutorial will take you through the most basic steps of working with *CLC Protein Workbench*. The tutorial introduces the user interface, shows how to create a folder, and demonstrates how to import your own existing data into the program.

When you open *CLC Protein Workbench* for the first time, the user interface looks like figure 2.1.

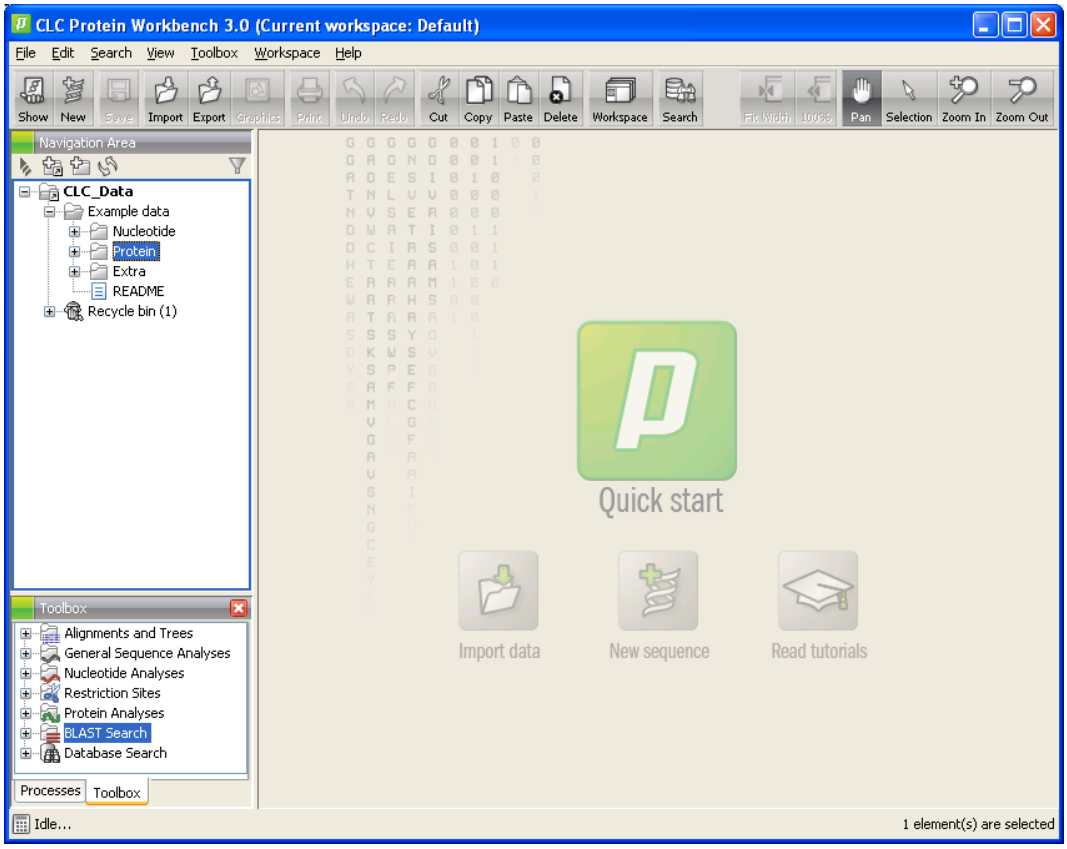


Figure 2.1: *The user interface as it looks when you start the program for the first time. (Windows version of* **CLC Protein Workbench***. The interface is similar for Mac and Linux.)*

At this stage, the important issues are the **Navigation Area** and the **View Area**.

The **Navigation Area** to the left is where you keep all your data for use in the program. Most analyses of *CLC Protein Workbench* require that the data is saved in the **Navigation Area**. There are several ways to get data into the **Navigation Area**, and this tutorial describes how to import existing data.

The **View Area** is the main area to the right. This is where the data can be 'viewed'. In general,

a **View** is a display of a piece of data, and the **View Area** can include several **Views**. The **Views** are represented by tabs, and can be organized e.g. by using 'drag and drop'.

2.1.1 Creating a a folder

When *CLC Protein Workbench* is started there is one element in the **Navigation Area** called **CLC_Data**[1]. This element is a **Location**. A location points to a folder on your computer where your data for use with *CLC Protein Workbench* is stored.

The data in the location can be organized into folders. Create a folder:

File | **New** | **Folder (** **)**
or **Ctrl + Shift + N (⌘ + Shift + N on Mac)**

Name the folder 'My folder' and press **Enter**.

2.1.2 Import data

Next, we want to import a sequence called HUMDINUC.fsa (FASTA format) from our own Desktop into the new 'My folder'. (This file is chosen for demonstration purposes only - you may have another file on your desktop, which you can use to follow this tutorial. You can import all kinds of files.)

In order to import the HUMDINUC.fsa file:

> **Select 'My folder'** | **Import (** **) in the Toolbar** | **navigate to HUMDINUC.fsa on the desktop** | **Select**

The sequence is imported into the folder that was selected in the **Navigation Area**, before you clicked **Import**. Double-click the sequence in the **Navigation Area** to view it. The final result looks like figure 2.2.

2.2 Tutorial: View sequence

This brief tutorial will take you through some different ways to display a sequence in the program. The tutorial introduces zooming on a sequence, dragging tabs, and opening selection in new view.

We will be working with the sequence called *pcDNA3-atp8a1* located in the 'Cloning' folder in the Example data. Double-click the sequence in the **Navigation Area** to open it. The sequence is displayed with annotations above it. (See figure 2.3).

As default, *CLC Protein Workbench* displays a sequence with annotations (colored arrows on the sequence like the green promoter region annotation in figure 2.3) and zoomed to see the residues.

In this tutorial we want to have an overview of the whole sequence. Hence;

> **click Zoom Out (** **) in the Toolbar** | **click the sequence until you can see the whole sequence**

[1] If you have downloaded the example data, this will be placed as a folder in *CLC_Data*

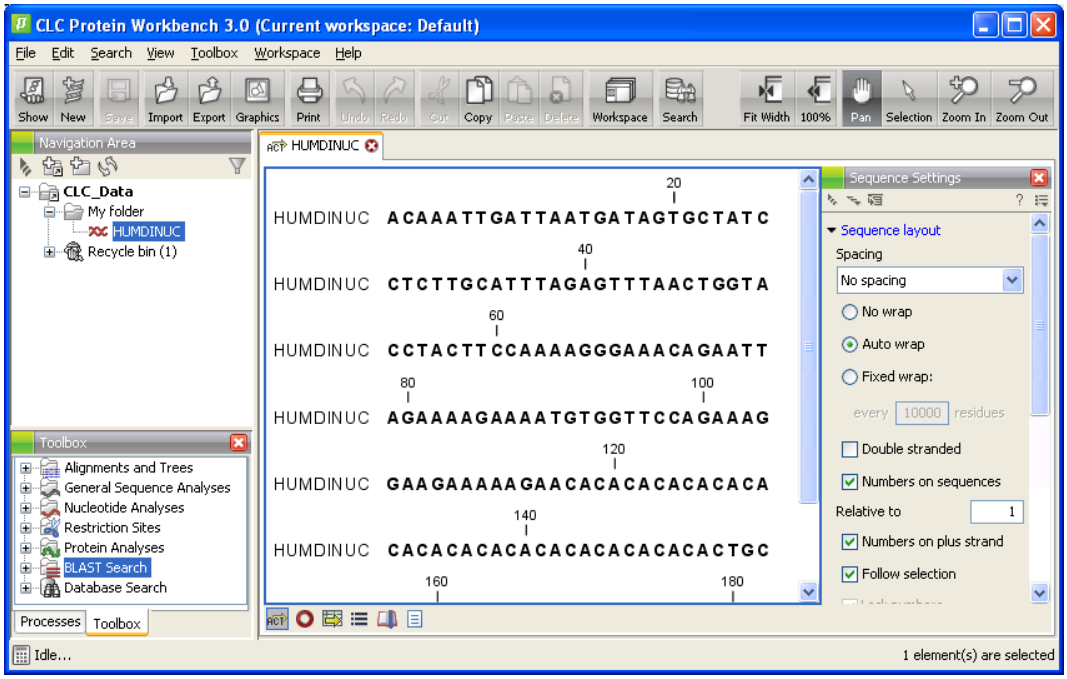


Figure 2.2: *The HUMDINUC file is imported and opened.*

Figure 2.3: *Sequence pcDNA3-atp8a1 opened in a view.*

This sequence is circular, which is indicated by << and >> at the beginning and the end of the sequence.

In the following we will show how the same sequence can be displayed in two different views - one linear view and one circular view. First, zoom in to see the residues again by using the **Zoom In** ( ) or the **100%** ( ). Then we make a split view by:

press and hold the Ctrl-button on the keyboard (⌘ on Mac) | click Show as Circular ( ) at the bottom of the view

This opens an additional view of the vector with a circular display, as can be seen in figure 2.4.

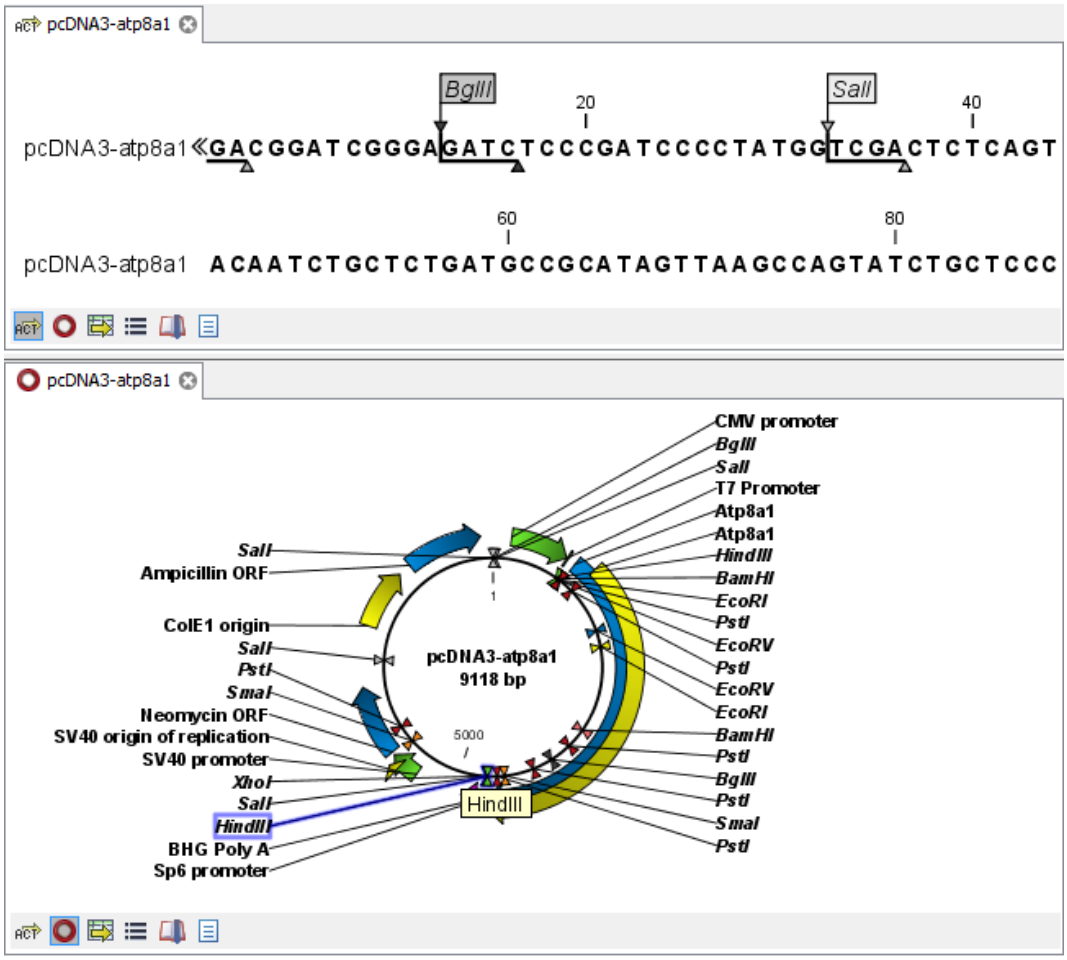


Figure 2.4: *The resulting two views which are split horizontally.*

Make a selection on the circular sequence (remember to switch to the **Selection** ( ) tool in the tool bar) and note that this selection is also reflected in the linear view above.

2.3 Tutorial: Side Panel Settings

This brief tutorial will show you how to use the **Side Panel** to change the way your sequences, alignments and other data are shown. You will also see how to save the changes that you made in the **Side Panel** .

Open the protein alignment located under *Protein orthologs* in the **Example data**. The initial view of the alignment has colored the residues according to the Rasmol color scheme, and the alignment is automatically wrapped to fit the width of the view (shown in figure 2.5).

Now, we are going to modify how this alignment is displayed. For this, we use the settings in the **Side Panel** to the right. All the settings are organized into groups, which can be expanded / collapsed by clicking the name of the group. The first group is **Sequence Layout** which is expanded by default.

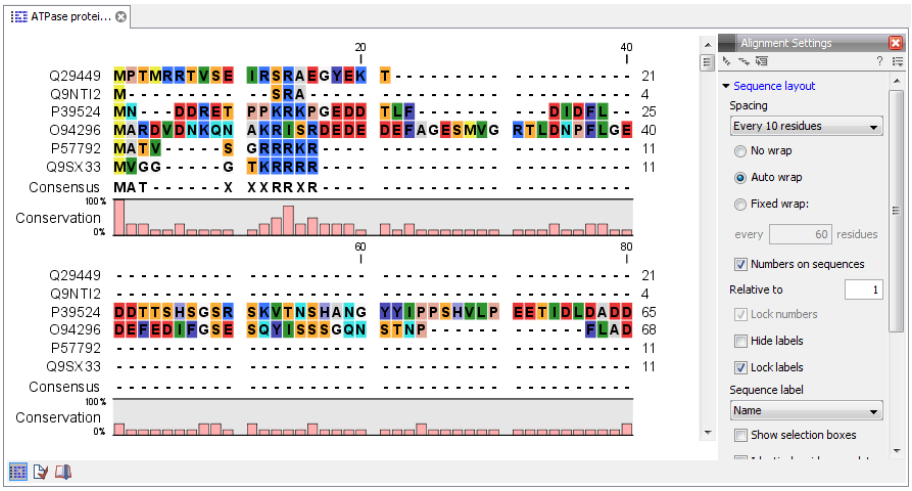


Figure 2.5: *The protein alignment as it looks when you open it with background color according to the Rasmol color scheme and automatically wrapped.*

First, select **No wrap** in the **Sequence Layout**. This means that each sequence in the alignment is kept on the same line. To see more of the alignment, you now have to scroll horizontally.

Next, expand the **Annotation Layout** group and select **Show Annotations**. Set the **Offset** to "More offset" and set the **Label** to "Stacked".

Expand the **Annotation Types** group. Here you will see a list of the types annotation that are carried by the sequences in the alignment (see figure 2.6).

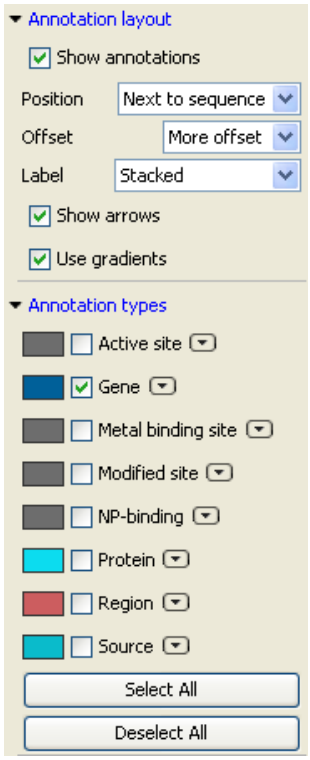


Figure 2.6: *The Annotation Layout and the Annotation Types in the Side Panel.*

Check the "Region" annotation type, and you will see the regions as red annotations on the

sequences.

Next, we will change the way the residues are colored. Click the **Alignment Info** group and under **Conservation**, check "Background color". This will use a gradient as background color for the residues. You can adjust the coloring by dragging the small arrows above the color box.

2.3.1 Saving the settings in the Side Panel

Now the alignment should look similar to figure 2.7.

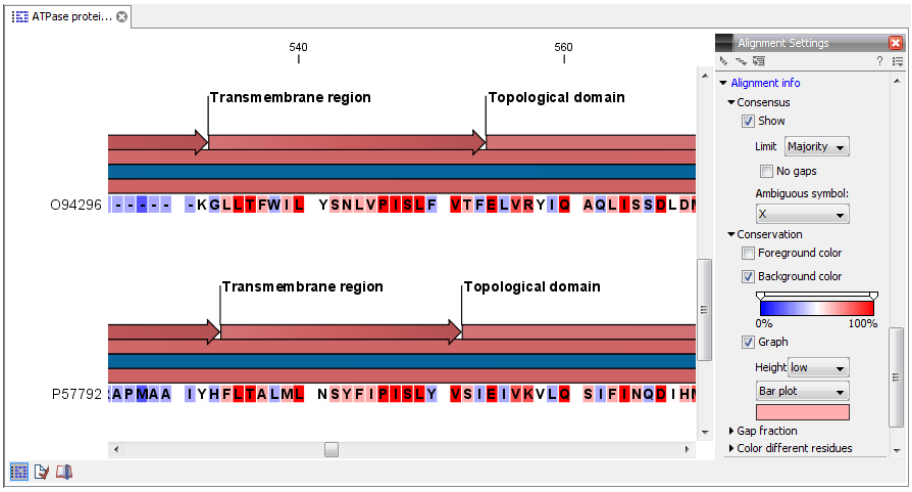


Figure 2.7: *The alignment when all the above settings have been changed.*

At this point, if you just close the view, the changes made to the **Side Panel** will not be saved. This means that you would have to perform the changes again next time you open the alignment. To save the changes to the **Side Panel**, click the **Save/Restore Settings** button ( ) at the top of the **Side Panel** and click **Save Settings** (see figure 2.8).

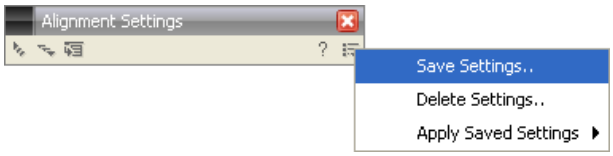


Figure 2.8: *Saving the settings of the Side Panel.*

This will open the dialog shown in figure 2.9.

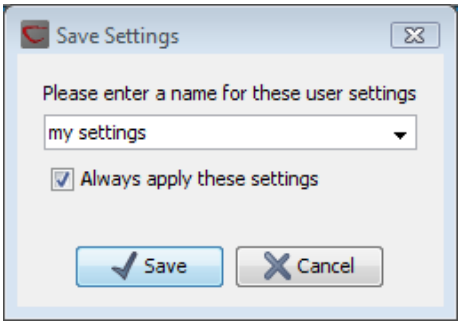


Figure 2.9: *Dialog for saving the settings of the Side Panel.*

In this way you can save the current state of the settings in the **Side Panel** so that you can apply

them to alignments later on. If you check **Always apply these settings**, these settings will be applied every time you open a view of the alignment.

Type "My settings" in the dialog and click **Save**.

2.3.2 Applying saved settings

When you click the **Save/Restore Settings** button ( ) again and select **Apply Saved Settings**, you will see "My settings" in the menu together with some pre-defined settings that the *CLC Protein Workbench* has created for you (see figure 2.10).

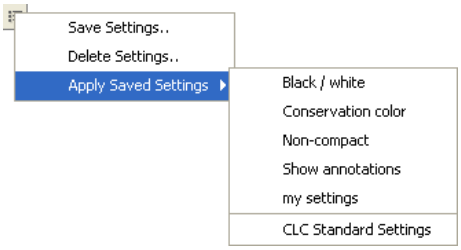


Figure 2.10: *Menu for applying saved settings.*

Whenever you open an alignment, you will be able to apply these settings. Each kind of view has its own list of settings that can be applied.

At the bottom of the list you will see the "CLC Standard Settings" which are the default settings for the view.

2.4 Tutorial: GenBank search and download

The *CLC Protein Workbench* allows you to search the NCBI GenBank database directly from the program, giving you the opportunity to both open, view, analyze and save the search results without using any other applications. To conduct a search in NCBI GenBank from *CLC Protein Workbench* you must be connected to the Internet.

This tutorial shows how to find a complete human hemoglobin DNA sequence in a situation where you do not know the accession number of the sequence.

To start the search:

Search | Search for Sequences at NCBI ( )

This opens the search view. We are searching for a DNA sequence, hence:

Nucleotide

Now we are going to adjust parameters for the search. By clicking **Add search parameters** you activate an additional set of fields where you can enter search criteria. Each search criterion consists of a drop down menu and a text field. In the drop down menu you choose which part of the NCBI database to search, and in the text field you enter what to search for:

Click Add search parameters until three search criteria are available | choose Organism in the first drop down menu | write 'human' in the adjoining text field | choose All Fields in the second drop down menu | write 'hemoglobin' in the adjoining text field | choose All Fields in the third drop down menu | write 'complete' in the adjoining text field

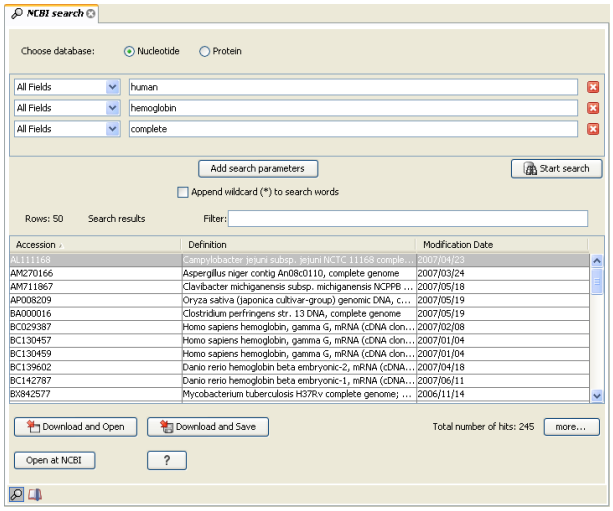

NCBI search
Choose database:
Nucleotide
Protein
All Fields
human
All Fields
hemoglobin
All Fields
complete
Add search parameters
Start search
Append wildcard (*) to search words
Rows: 50
Search results
Filter:
Accession
Definition
Modification Date
AL111168 Campylobacter jejuni subsp. jejuni NCTC 11168 comple... 2007/04/23
AM270166 Aspergillus niger contig An08c0110, complete genome 2007/03/24
AM711867 Clavibacter michiganensis subsp. michiganensis NCPPB ... 2007/05/18
AP008209 Oryza sativa (japonica cultivar-group) genomic DNA, c... 2007/05/19
BA000016 Clostridium perfringens str. 13 DNA, complete genome 2007/05/19
BC029387 Homo sapiens hemoglobin, gamma G, mRNA (cDNA clon... 2007/02/08
BC130457 Homo sapiens hemoglobin, gamma G, mRNA (cDNA clon... 2007/01/04
BC130459 Homo sapiens hemoglobin, gamma G, mRNA (cDNA clon... 2007/01/04
BC139602 Danio rerio hemoglobin beta embryonic-2, mRNA (cDNA... 2007/04/18
BC142787 Danio rerio hemoglobin beta embryonic-1, mRNA (cDNA... 2007/06/11
BX842577 Mycobacterium tuberculosis H37Rv complete genome; ... 2006/11/14
Download and Open
Download and Save
Total number of hits: 245
more...
Open at NCBI
?

Figure 2.11: *NCBI search view.*

Click **Start search ( )** to commence the search in NCBI.

2.4.1 Searching for matching objects

When the search is complete, the list of hits is shown. If the desired complete human hemoglobin DNA sequence is found, the sequence can be viewed by double-clicking it in the list of hits from the search. If the desired sequence is not shown, you can click the 'More' button below the list to see more hits.

2.4.2 Saving the sequence

The sequences which are found during the search can be displayed by double-clicking in the list of hits. However, this does not save the sequence. You can save one or more sequence by selecting them and:

click Download and Save

or **drag the sequences into the Navigation Area**

2.5 Tutorial: BLAST search

BLAST is an invaluable tool in bioinformatics. It has become central to identification of homologues and similar sequences, and can also be used for many other different purposes. This tutorial takes you through the steps of running a blast search in CLC Workbenches. If you plan to use blast for your research, we highly recommend that you read further about it. Understanding how blast works is key to setting up meaningful and efficient searches.

Suppose you are working with the ATP8a1 protein sequence which is a phospholipid-transporting ATPase expressed in the adult house mouse, *Mus musculus*. To obtain more information about this molecule you wish to query the peptides held in the Swiss-Prot* database to find homologous proteins in humans *Homo sapiens*, using the **Basic Local Alignment Search Tool** (BLAST) algorithm.

This tutorial involves running BLAST remotely using databases housed at the NCBI. Your computer must be connected to the internet to complete this tutorial.

2.5.1 Performing the BLAST search

Start out by:

select protein ATP8a1 | Toolbox | BLAST Search ( ) | NCBI BLAST ( )

In **Step 1** you can choose which sequence to use as query sequence. Since you have already chosen the sequence it is displayed in the **Selected Elements** list.

Click **Next**.

In **Step 2** (figure 2.12), choose the default BLAST program: **blastp: Protein sequence and database** and select the **Swiss-Prot** database in the **Database** drop down menu.

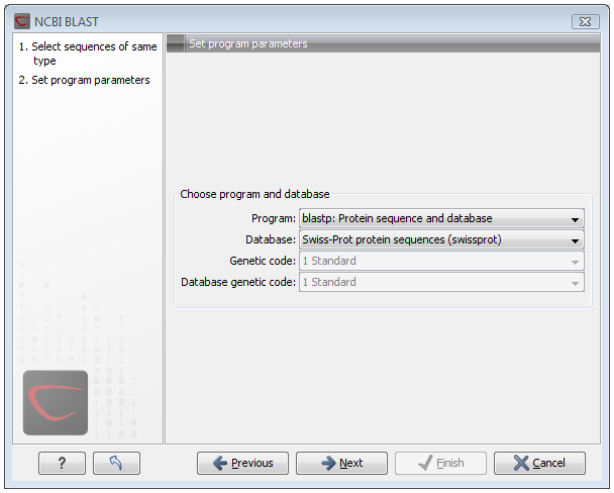


Figure 2.12: *Choosing BLAST program and database.*

Click **Next**.

In the **Limit by Entrez query** in **Step 3**, choose **Homo sapiens[ORGN]** from the drop down menu to arrive at the search configuration seen in figure 2.13. Including this term limits the query to proteins of human origin.

Choose to **Open** your results.

Click **Finish** to accept the parameter settings and begin the BLAST search.

The computer now contacts NCBI and places your query in the BLAST search queue. After a short while the result should be received and opened in a new view.

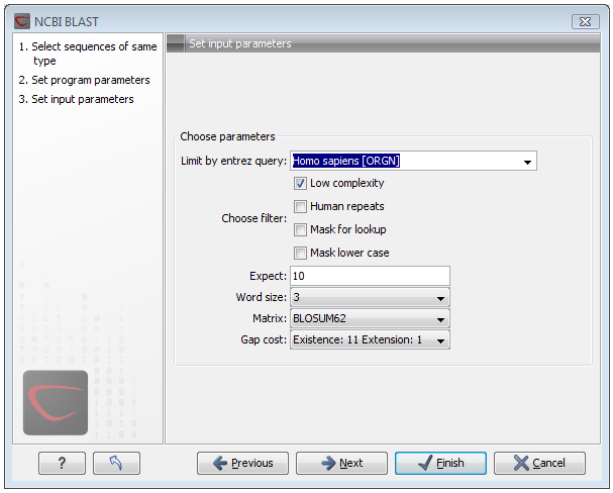


Figure 2.13: *The BLAST search is limited to homo sapiens[ORGN]. The remaining parameters are left as default.*

2.5.2 Inspecting the results

The output is shown in figure 2.14 and consists of a list of potential homologs that are sorted by their BLAST match-score and shown in descending order below the query sequence.

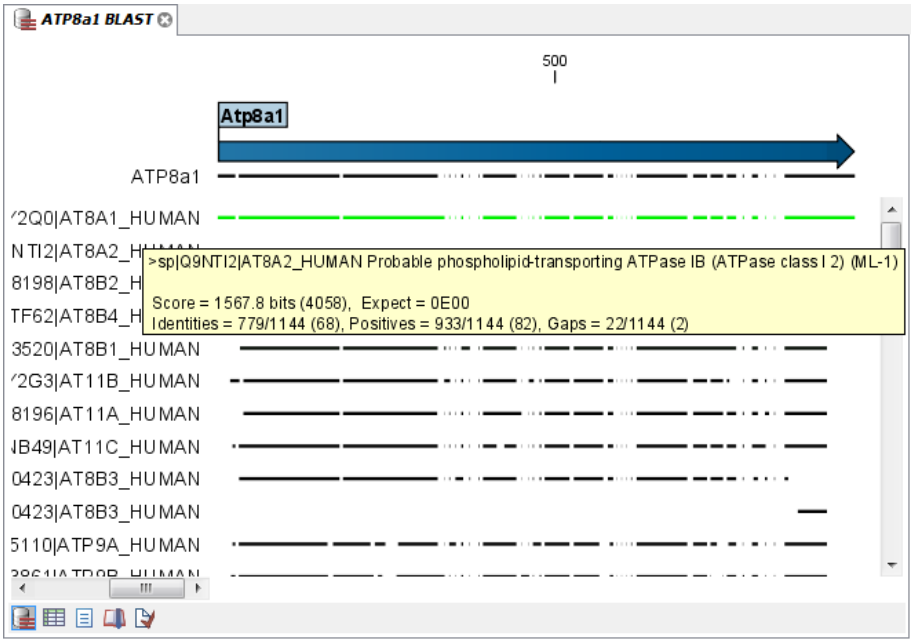

ATP8a1 BLAST
500
Atp8a1
ATP8a1
>sp|Q9NTI2|AT8A2_HUMAN Probable phospholipid-transporting ATPase IB (ATPase class I 2) (ML-1)
Score = 1567.8 bits (4058), Expect = 0E00
Identities = 779/1144 (68), Positives = 933/1144 (82), Gaps = 22/1144 (2)

Figure 2.14: *Output of a BLAST search. By holding the mouse pointer over the lines you can get information about the sequence.*

Try placing your mouse cursor over a potential homologous sequence. You will see that a context box appears containing information about the sequence and the match-scores obtained from the BLAST algorithm.

The lines in the BLAST view are the actual sequences which are downloaded. This means that you can zoom in and see the actual alignment:

> **Zoom in in the Tool Bar ( ) | Click in the BLAST view a number of times until you see the residues**

Now we will focus our attention on sequence Q9Y2Q0 - the BLAST hit that is at the top of the list. To download the full sequence:

right-click the line representing sequence Q9Y2Q0 | Download Full Hit Sequence from NCBI

This opens the sequence. However, the sequence is not saved yet. Drag and drop the sequence into the **Navigation Area** to save it. This homologous sequence is now stored in the *CLC Protein Workbench* and you can use it to gain information about the query sequence by using the various tools of the workbench, e.g. by studying its textual information, by studying its annotation or by aligning it to the query sequence.

2.5.3 Using the BLAST table view

As an alternative to the graphic BLAST view, you can click the Table View ( ) at the bottom. This will display a tabular view of the BLASt hits as shown in figure 2.15.

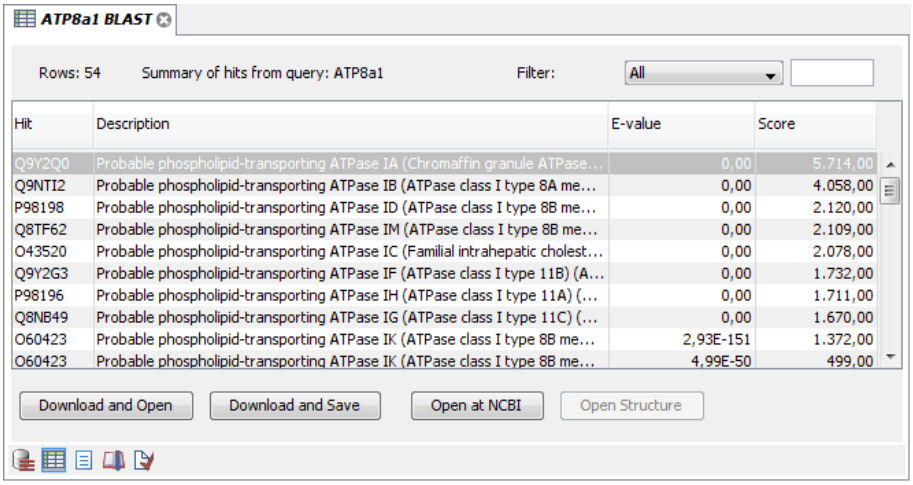

ATP8a1 BLAST
Rows: 54
Summary of hits from query: ATP8a1
Filter:
All
Hit
Description
E-value
Score
Q9Y2Q0 Probable phospholipid-transporting ATPase IA (Chromaffin granule ATPase... 0,00 5.714,00
Q9NTI2 Probable phospholipid-transporting ATPase IB (ATPase class I type 8A me... 0,00 4.058,00
P98198 Probable phospholipid-transporting ATPase ID (ATPase class I type 8B me... 0,00 2.120,00
Q8TF62 Probable phospholipid-transporting ATPase IM (ATPase class I type 8B me... 0,00 2.109,00
O43520 Probable phospholipid-transporting ATPase IC (Familial intrahepatic cholest... 0,00 2.078,00
Q9Y2G3 Probable phospholipid-transporting ATPase IF (ATPase class I type 11B) (A... 0,00 1.732,00
P98196 Probable phospholipid-transporting ATPase IH (ATPase class I type 11A) (... 0,00 1.711,00
Q8NB49 Probable phospholipid-transporting ATPase IG (ATPase class I type 11C) (... 0,00 1.670,00
O60423 Probable phospholipid-transporting ATPase IK (ATPase class I type 8B me... 2,93E-151 1.372,00
O60423 Probable phospholipid-transporting ATPase IK (ATPase class I type 8B me... 4,99E-50 499,00
Download and Open
Download and Save
Open at NCBI
Open Structure

Figure 2.15: *Output of a BLAST search shown in a table.*

This view provides more statistics about the hits, and you can use the filter to search for e.g. a specific type of protein etc. If you wish to download several of the hit sequences, this is easily done in this view. Simply select the relevant sequences and drag them into a folder in the **Navigation Area**.

2.6 Tutorial: Tips for specialized BLAST searches

Here, you will learn how to:

- Use BLAST to find the gene coding for a protein in a genomic sequence.
- Find primer binding sites on genomic sequences
- Identify remote protein homologues.

Following through these sections of the tutorial requires some experience using the Workbench, so if you get stuck at some point, we recommend going through the more basic tutorials first.

2.6.1 Locate a protein sequence on the chromosome

If you have a protein sequence but want to see the actual location on the chromosome this is easy to do using BLAST.

In this example we wish to map the protein sequence of the Human beta-globin protein to a chromosome. We know in advance that the beta-globin is located somewhere on chromosome 11.

Data used in this example can be downloaded from GenBank:

Search | **Search for Sequences at NCBI (** **)**

Human chromosome 11 (NC_000011) consists of 134452384 nucleotides and the beta-globin (AAA16334) protein has 147 amino acids.

BLAST configuration

Next, conduct a local BLAST search:

Toolbox | **BLAST Search (** **)** | **Local BLAST (** **)**

Select the protein sequence as query sequence and click **Next**. Since you wish to BLAST a protein sequence against a nucleotide sequence, use **tblastn** which will automatically translate the nucleotide sequence selected as database.

As **Target** select *NC_000011* that you downloaded. If you are used to BLAST, you will know that you usually have to create a BLAST database before BLASTing, but the Workbench does this "on the fly" when you just select one or more sequences.

Click **Next**, leave the parameters at their default, click **Next** again, and then **Finish**.

Inspect BLAST result

When the BLAST result appears make a split view so that both the table and graphical view is visible (see figure 2.16). This is done by pressing Ctrl (⌘ on Mac) while clicking the table view ( ) at the bottom of the view.

In the table start out by showing two additional columns; "% Positive" and "Query start". These should simply be checked in the Side Panel.

Now, sort the BLAST table view by clicking the column header "% Positive". Then, press and hold the Ctrl button (⌘ on Mac) and click the header "Query start". Now you have sorted the table first on % Positive hits and then the start position of the query sequence. Now you see that you actually have three regions with a 100% positive hit but at different locations on the chromosome sequence (see figure 2.16).

Why did we find, on the protein level, three identical regions between our query protein sequence and nucleotide database?

The beta-globin gene is known to have three exons and this is exactly what we find in the BLAST search. Each translated exon will hit the corresponding sequence on the chromosome.

If you place the mouse cursor on the sequence hits in the graphical view, you can see the reading frame which is -1, -2 and -3 for the three hits, respectively.

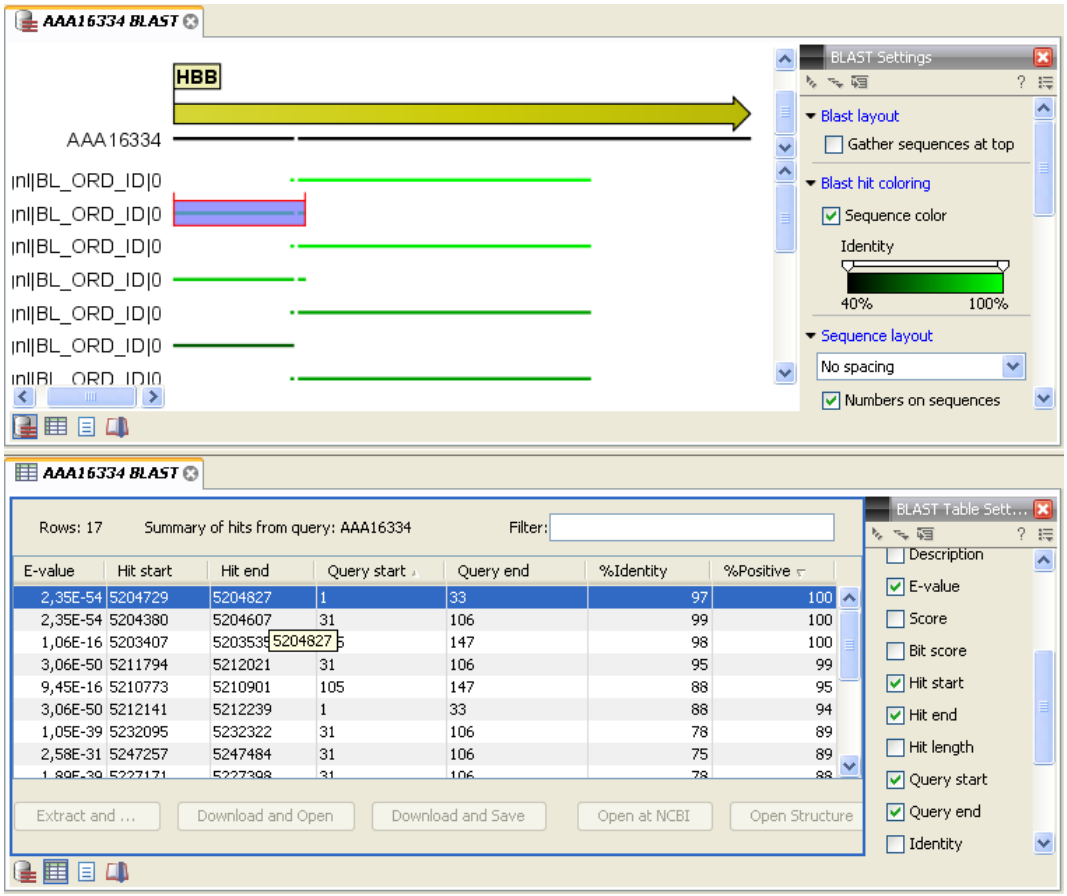


Figure 2.16: *Placement of translated nucleotide sequence hits on the Human beta-globin.*

Verify the result

Open NC_000011 in a view, and go to the Hit start position (5,204,729) and zoom to see the blue gene annotation. You can now see the exon structure of the Human beta-globin gene showing the three exons on the reverse strand (see figure 2.17).

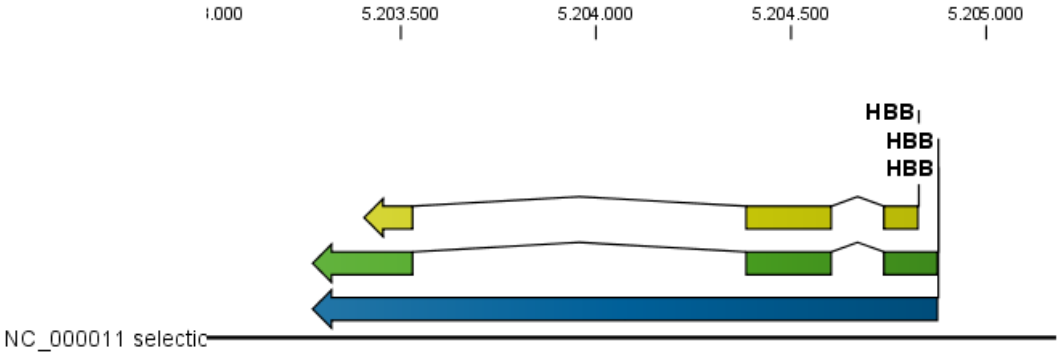


Figure 2.17: *Human beta-globin exon view.*

If you wish to verify the result, make a selection covering the gene region and open it in a new view:

right-click | **Open Selection in New View (** **)** | **Save (** **)**

Save the sequence, and perform a new BLAST search:

- Use the new sequence as query.
- Use BLASTx
- Use the protein sequence, AAA16334, as database

Using the genomic sequence as query, the mapping of the protein sequence to the exons is visually very clear as shown in figure 2.18.

Figure 2.18: *Verification of the result: at the top a view of the whole BLAST result. At the bottom the same view is zoomed in on exon 3 to show the amino acids.*

In theory you could use the chromosome sequence as query, but the performance would not be optimal: it would take a long time, and the computer might run out of memory.

In this example, you have used well-annotated sequences where you could have searched for the name of the gene instead of using BLAST. However, there are other situations where you either do not know the name of the gene, or the genomic sequence is poorly annotated. In these cases, the approach described in this tutorial can be very productive.

2.6.2 BLAST for primer binding sites

You can adjust the BLAST parameters so it becomes possible to match short primer sequences against a larger sequence. Then it is easy to examine whether already existing lab primers can

be reused for other purposes, or if the primers you designed are specific.

Purpose	Program	Word size	Low complexity filter	Expect value
Standard BLAST	blastn	11	On	10
Primer search	blastn	7	Off	1000

These settings are shown in figure 2.19.

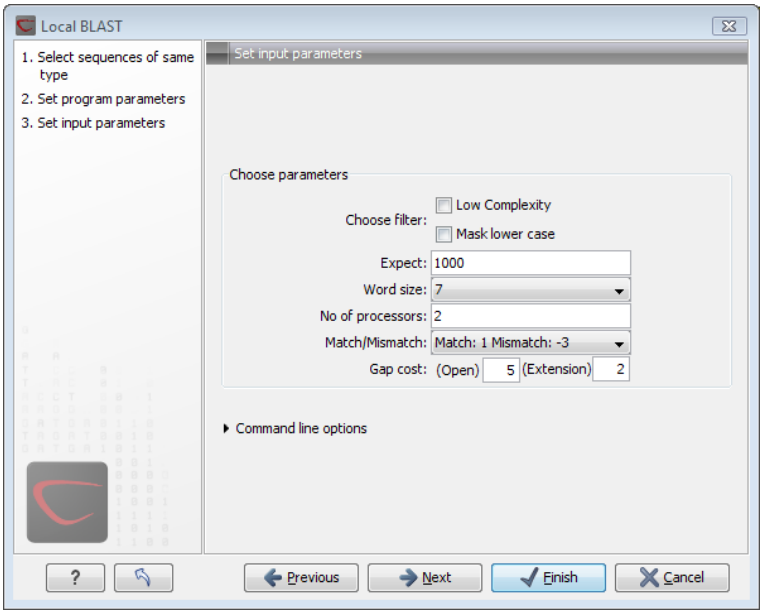


Figure 2.19: *Settings for searching for primer binding sites.*

2.6.3 Finding remote protein homologues

If you look for short identical peptide sequences in a database, the standard BLAST parameters will have to be reconfigured. Using the parameters described below, you are likely to be able to identify whether antigenic determinants will cross react to other proteins.

Purpose	Program	Word size	Low complexity filter	Expect value	Scoring matrix
Standard BLAST	blastp	3	On	10	BLSUM62
Remote homologues	blastp	2	Off	20000	PAM30

These settings are shown in figure 2.20.

2.6.4 Further reading

A valuable source of information about BLAST can be found at http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=ProgSelectionGuide.

Remember that BLAST is a heuristic method. This means that certain assumptions are made to allow searches to be done in a reasonable amount of time. Thus you cannot trust BLAST search results to be accurate. For very accurate results you should consider using other algorithms, such as Smith-Waterman. You can read "Bioinformatics explained: BLAST versus Smith-Waterman" here: http://www.clcbio.com/BE.

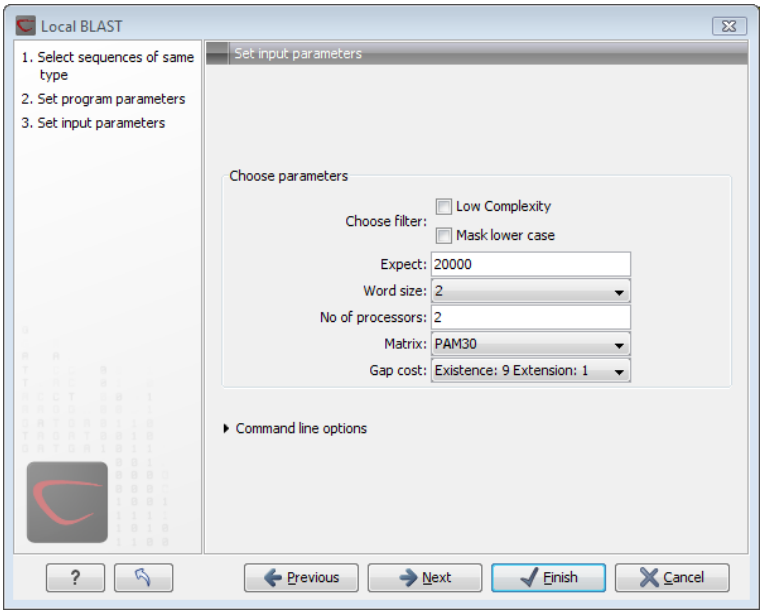


Figure 2.20: *Settings for searching for remote homologues.*

2.7 Tutorial: Proteolytic cleavage detection

This tutorial shows you how to find cut sites and see an overview of fragments when cleaving proteins with proteolytic cleavage enzymes.

Suppose you are working with protein ATP8a1 from the example data, and you wish to see where the enzyme *trypsin* will cleave the protein. Furthermore, you want to see details for the resulting fragments which are between 10 and 15 amino acids long.

select protein ATP8a1 | Toolbox | Protein Analyses ( ) | Proteolytic Cleavage

This opens **Step 1** of the Proteolytic Cleavage dialog. In this step you can choose which sequences to include in the analysis. Since you have already chosen ATP8a1, click **Next**.

In this step you should select **Trypsin**. This is illustrated in figure 2.21.

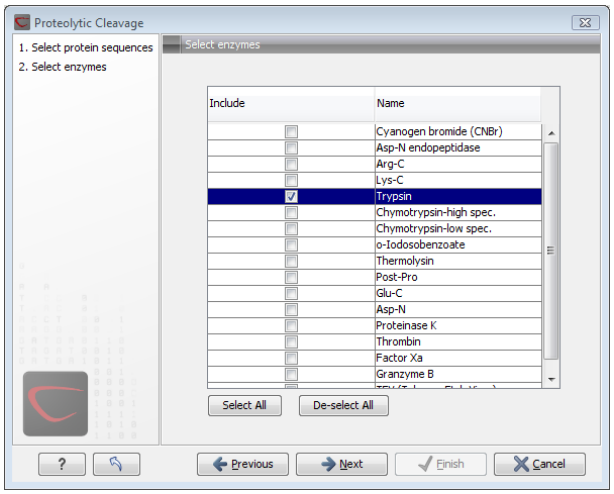


Figure 2.21: *Selecting trypsin as the cleaving enzyme.*

Click **Next** to go to **Step 3** of the dialog.

In **Step 3** you can adjust the parameters for which fragments of the cleavage you want to include in the table output of the analysis.

> **Type '10' in the Min. fragment length | Check the box: Max. fragment length | enter '15' in the corresponding text field**

These parameter adjustments are shown in figure 2.22:

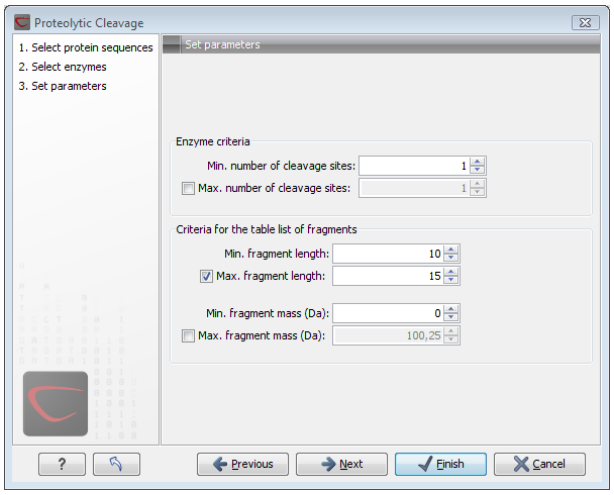


Figure 2.22: *Adjusting the output from the cleavage to include fragments which are between 10 and 15 amino acids long.*

Click **Finish** to make the analysis. The result of the analysis can be seen in figure 2.23

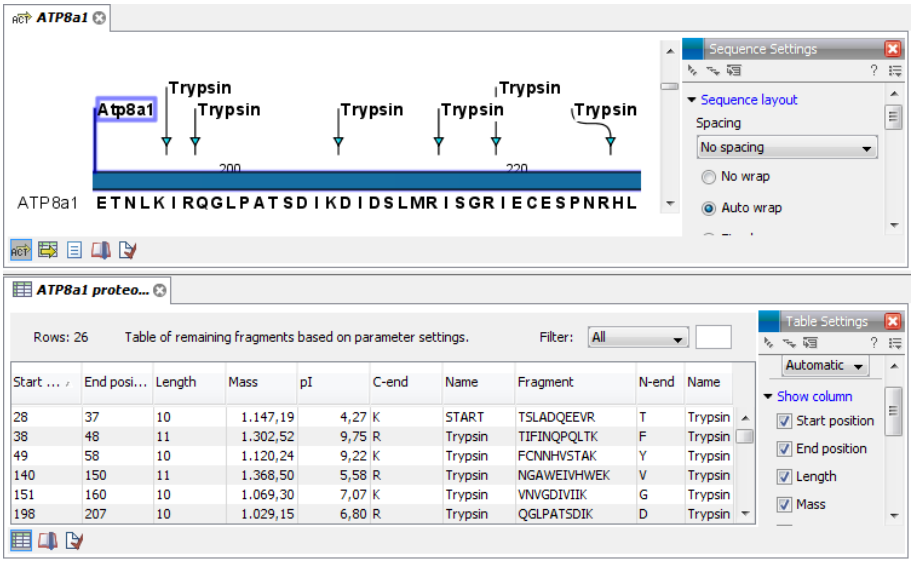


Figure 2.23: *The output of the proteolytic cleavage shows the cleavage sites as annontations in the protein sequence. The accompanying table lists all the fragments which are between 10 and 15 amino acids long.*

Note! The output of proteolytic cleavage is two related views. The sequence view displays annotations where the sequence is cleaved. The table view shows information about the fragments satisfying the parameters set in the dialog. Subsequently, if you have restricted the fragment parameters, you might have more annotations on the sequence than fragments in the

table.

If you conduct another proteolytic cleavage on the same sequence, the output consists of: (possibly) new annotations on the original sequence and an additional table view, listing all fragments.

2.8 Tutorial: Align protein sequences

This tutorial outlines some of the alignment functionality of the *CLC Protein Workbench*. In addition to creating alignments of nucleotide or peptide sequences, the software offers several ways to view alignments. The alignments can then be used for building phylogenetic trees.

Sequences must be available via the **Navigation Area** to be included in an alignment. If you have sequences open in a View that you have not saved, then you just need to select the view tab and press Ctrl + S (or ⌘ + S on Mac) to save them.

In this tutorial six protein sequences from the Example data folder will be aligned. (See figure 2.24).

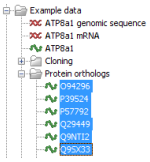


Figure 2.24: *Six protein sequences in 'Sequences' from the 'Protein orthologs' folder of the Example data.*

To align the sequences:

> **select the sequences from the 'Protein' folder under 'Sequences'** | **Toolbox** | **Alignments and Trees ()** | **Create Alignment ()**

2.8.1 The alignment dialog

This opens the dialog shown in figure 2.25.

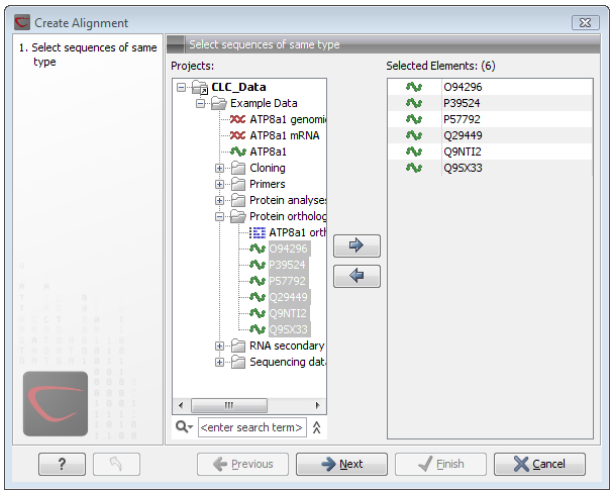


Figure 2.25: *The alignment dialog displaying the six protein sequences.*

It is possible to add and remove sequences from **Selected Elements** list. Since we had already selected the eight proteins, just click **Next** to adjust parameters for the alignment.

Clicking **Next** opens the dialog shown in figure 2.26.

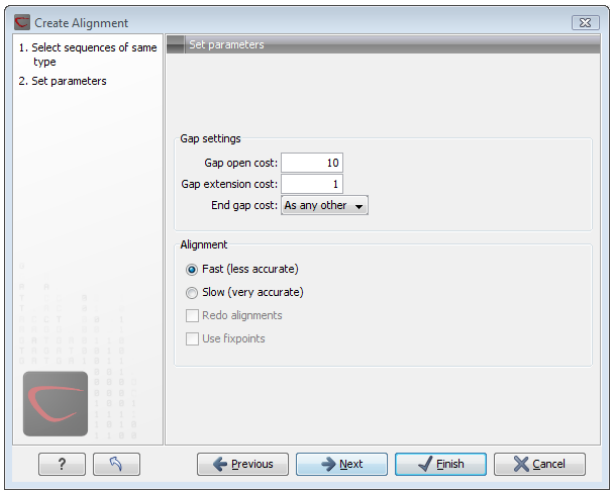


Figure 2.26: *The alignment dialog displaying the available parameters which can be adjusted.*

Leave the parameters at their default settings. An explanation of the parameters can be found by clicking the help button (?). Alternatively, a tooltip is displayed by holding the mouse cursor on the parameters.

Click **Finish** to start the alignment process which is shown in the **Toolbox** under the **Processes** tab. When the program is finished calculating it displays the alignment (see fig. 2.27):

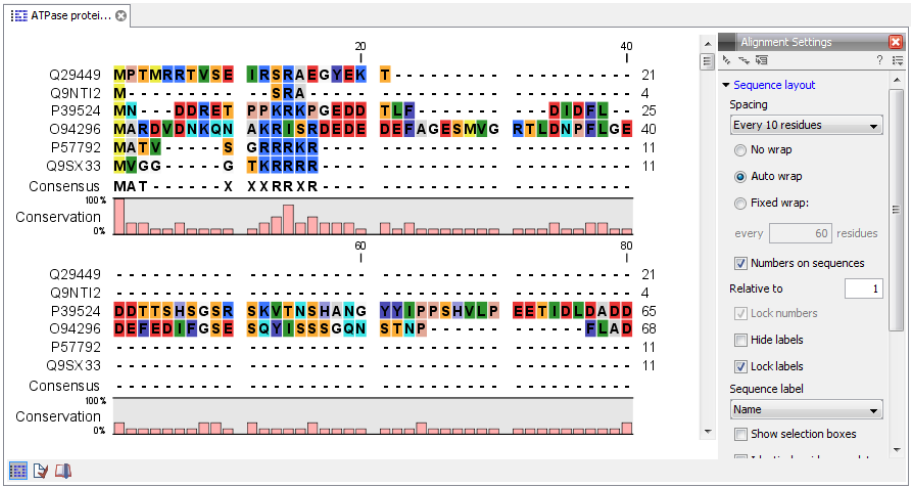


Figure 2.27: *The resulting alignment.*

Note! The new alignment is not saved automatically.

To save the alignment, drag the tab of the alignment view into the **Navigation Area**.

Installing the Additional Alignments plugin gives you access to other alignment algorithms: ClustalW (Windows/Mac/Linux), Muscle (Windows/Mac/Linux), T-Coffee (Mac/Linux), MAFFT (Mac/Linux), and Kalign (Mac/Linux). The Additional Alignments Module can be downloaded from `http://www.clcbio.com/plugins`. Note that you will need administrative privileges on your

system to install it.

2.9 Tutorial: Create and modify a phylogenetic tree

You can make a phylogenetic tree from an existing alignment. (See how to create an alignment in the tutorial: "Align protein sequences").

We use the 'ATPase protein alignment' located in 'Protein orthologs' in the Example data. To create a phylogenetic tree:

> **click the 'ATPase protein alignment' in the Navigation Area | Toolbox | Alignments and Trees ( ) | Create Tree (:)**

A dialog opens where you can confirm your selection of the alignment. Click **Next** to move to the next step in the dialog where you can choose between the neighbor joining and the UPGMA algorithms for making trees. You also have the option of including a bootstrap analysis of the result. Leave the parameters at their default, and click **Finish** to start the calculation, which can be seen in the **Toolbox** under the **Processes** tab. After a short while a tree appears in the **View Area** (figure 2.28).

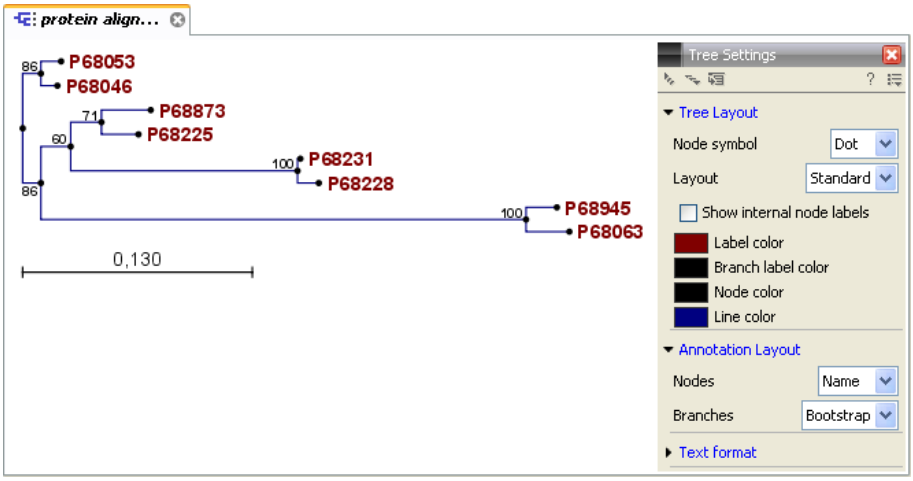


Figure 2.28: *After choosing which algorithm should be used, the tree appears in the View Area. The Side panel in the right side of the view allows you to adjust the way the tree is displayed.*

2.9.1 Tree layout

Using the **Side Panel** (in the right side of the view), you can change the way the tree is displayed.

Click **Tree Layout** and open the **Layout** drop down menu. Here you can choose between standard and topology layout. The topology layout can help to give an overview of the tree if some of the branches are very short.

When the sequences include the appropriate annotation, it is possible to choose between the accession number and the species names at the leaves of the tree. Sequences downloaded from GenBank, for example, have this information. The **Labels** preferences allows these different node annotations as well as different annotation on the branches.

The branch annotation includes the bootstrap value, if this was selected when the tree was calculated. It is also possible to annotate the branches with their lengths.

2.10 Tutorial: Find restriction sites

This tutorial will show you how to find restriction sites and annotate them on a sequence.

There are two ways of finding and showing restriction sites. In many cases, the dynamic restriction sites found in the **Side Panel** of sequence views will be useful, since it is a quick and easy way of showing restriction sites. In the **Toolbox** you will find the other way of doing restriction site analyses. This way provides more control of the analysis and gives you more output options, e.g. a table of restriction sites and a list of restriction enzymes that can be saved for later use. In this tutorial, the first section describes how to use the Side Panel to show restriction sites, whereas the second section describes the restriction map analysis performed from the **Toolbox**.

2.10.1 The Side Panel way of finding restriction sites

When you open a sequence, there is a **Restriction sites** setting in the **Side Panel**. By default, 10 of the most popular restriction enzymes are shown (see figure 2.29).

Figure 2.29: *Showing restriction sites of ten restriction enzymes.*

The restriction sites are shown on the sequence with an indication of cut site and recognition sequence. In the list of enzymes in the **Side Panel**, the number of cut sites is shown in parentheses for each enzyme (e.g. *SalI* cuts three times). If you wish to see the recognition sequence of the enzyme, place your mouse cursor on the enzyme in the list for a short moment, and a tool tip will appear.

You can add or remove enzymes from the list by clicking the **Manage enzymes** button.

2.10.2 The Toolbox way of finding restriction sites

Suppose you are working with sequence 'ATP8a1 mRNA' from the example data, and you wish to know which restriction enzymes will cut this sequence exactly once and create a 3' overhang. Do the following:

> **select the ATP8a1 mRNA | Toolbox in the Menu Bar | Restriction Sites ( ) | Restriction Site Analysis ( )**

Click **Next** to set parameters for the restriction map analysis.

In this step first select **Use existing enzyme list** and click the **Browse for enzyme list** button ( ). Select the 'Popular enzymes' in the Cloning folder under Enzyme lists.

Then write 3' into the filter below to the left. Select all the enzymes and click the **Add** button ( ). The result should be like in figure 2.30.

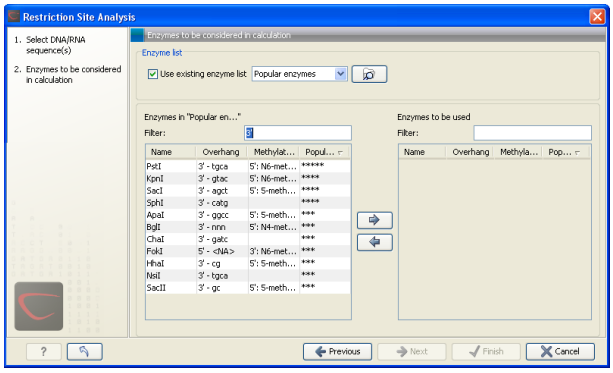


Figure 2.30: *Selecting enzymes.*

Click **Next**. In this step you specify that you want to show enzymes that cut the sequence only once. This means that you should de-select the **Two restriction sites** checkbox.

Click **Next** and select that you want to **Add restriction sites as annotations on sequence** and **Create restriction map**. (See figure 2.31).

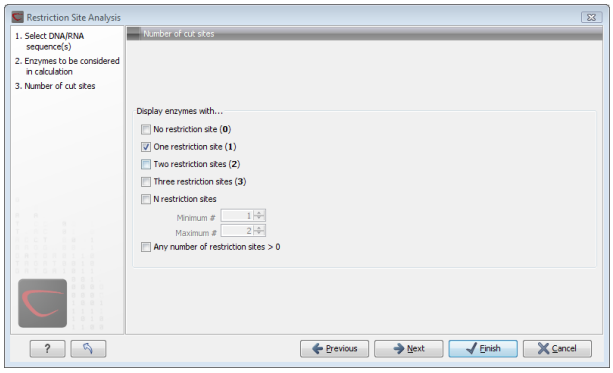


Figure 2.31: *Selecting output for restriction map analysis.*

Click **Finish** to start the restriction map analysis.

View restriction site

The restriction sites are shown in two views: one view is in a tabular format and the other view displays the sites as annotations on the sequence.

The result is shown in figure 2.32. The restriction map at the bottom can also be shown as a

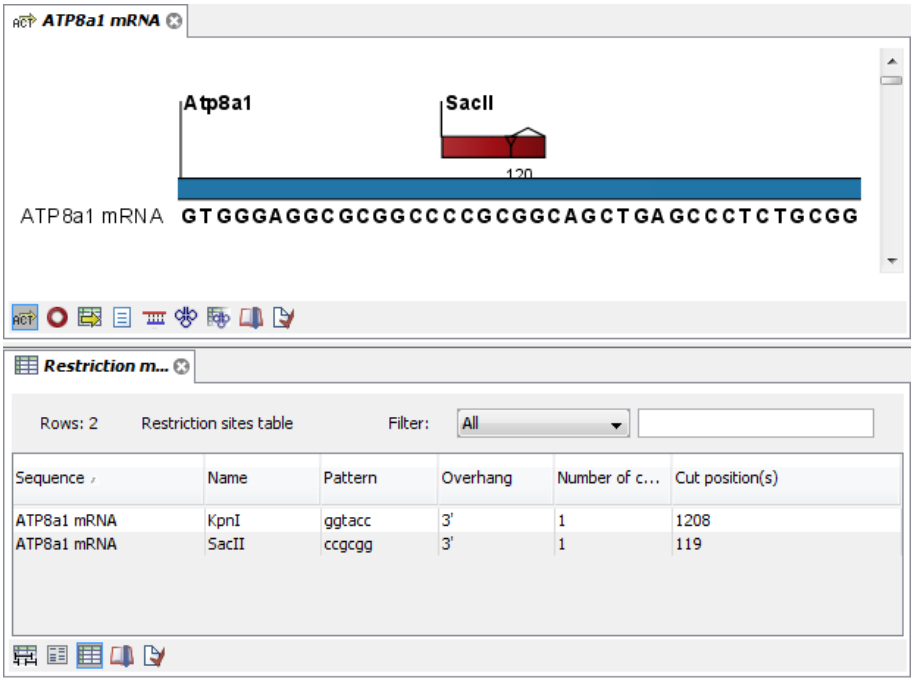


Figure 2.32: *The result of the restriction map analysis is displayed in a table at the bottom and as annotations on the sequence in the view at the top.*

table of fragments produced by cutting the sequence with the enzymes:

Click the Fragments button ( ) at the bottom of the view

In a similar way the fragments can be shown on a virtual gel:

Click the Gel button ( ) at the bottom of the view

Part II

Core Functionalities

Chapter 3

User interface

Contents

3.1 Navigation Area 62
3.1.1 Data structure 63
3.1.2 Create new folders 65
3.1.3 Sorting folders 65
3.1.4 Multiselecting elements 65
3.1.5 Moving and copying elements 65
3.1.6 Change element names 67
3.1.7 Delete elements 68
3.1.8 Show folder elements in a table 68
3.2 View Area 69
3.2.1 Open view 70
3.2.2 Show element in another view 71
3.2.3 Close views 71
3.2.4 Save changes in a view 72
3.2.5 Undo/Redo 72
3.2.6 Arrange views in View Area 73
3.2.7 Side Panel 75
3.3 Zoom and selection in View Area 76
3.3.1 Zoom In 76
3.3.2 Zoom Out 76
3.3.3 Fit Width 77
3.3.4 Zoom to 100% 77
3.3.5 Move 77
3.3.6 Selection 77
3.3.7 Changing compactness 77
3.4 Toolbox and Status Bar 78
3.4.1 Processes 78
3.4.2 Toolbox 79
3.4.3 Status Bar 79
3.5 Workspace 79

3.5.1 Create Workspace 79
3.5.2 Select Workspace 79
3.5.3 Delete Workspace 80
3.6 List of shortcuts 80

This chapter provides an overview of the different areas in the user interface of *CLC Protein Workbench*. As can be seen from figure 3.1 this includes a **Navigation Area**, **View Area**, **Menu Bar**, **Toolbar**, **Status Bar** and **Toolbox**.

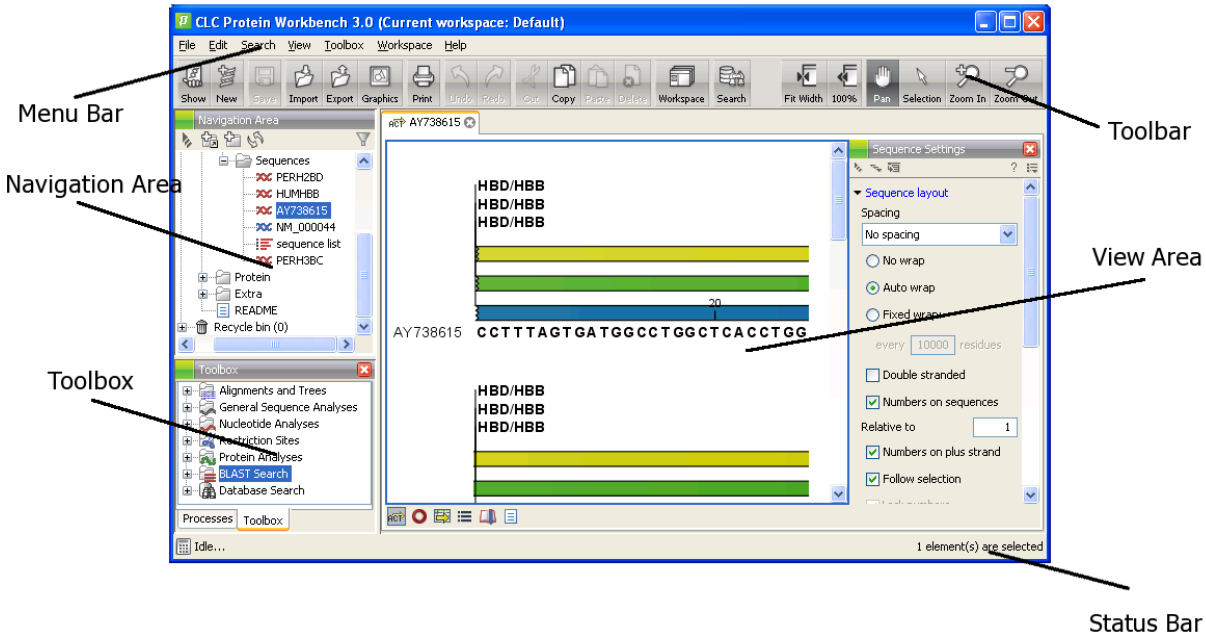


Figure 3.1: *The user interface consists of the Menu Bar, Toolbar, Status Bar, Navigation Area, Toolbox, and View Area.*

3.1 Navigation Area

The **Navigation Area** is located in the left side of the screen, under the **Toolbar** (see figure 3.2). It is used for organizing and navigating data. Its behavior is similar to the way files and folders are usually displayed on your computer.

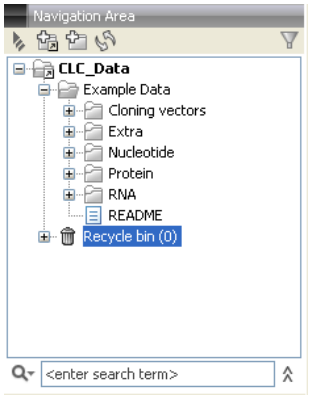


Figure 3.2: *The Navigation Area.*

3.1.1 Data structure

The data in the **Navigation Area** is organized into a number of **Locations**. When the *CLC Protein Workbench* is started for the first time, there is one location called *CLC_Data* (unless your computer administrator has configured the installation otherwise).

A location represents a folder on the computer: The data shown under a location in the **Navigation Area** is stored on the computer in the folder which the location points to.

This is explained visually in figure 3.3.

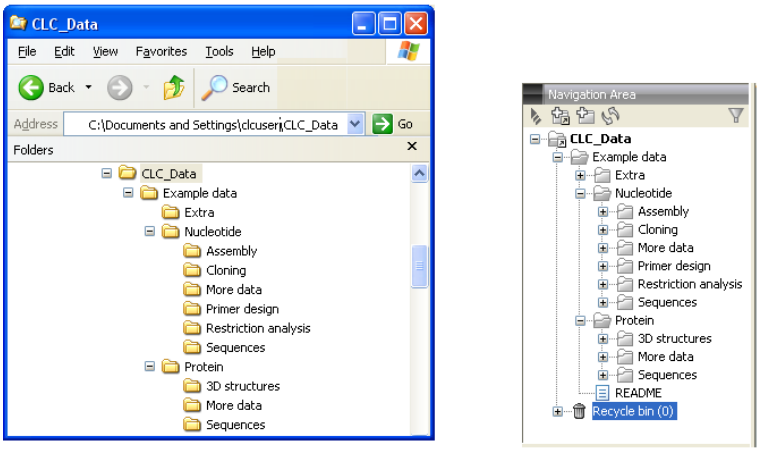


Figure 3.3: *In this example the location called 'CLC_Data' points to the folder at C:\Documents and settings\clcuser\CLC_Data.*

Adding locations

Per default, there is one location in the **Navigation Area** called CLC_Data. It points to the following folder:

- On Windows: C:\Documents and settings\<username>\CLC_Data
- On Mac: ~/CLC_Data
- On Linux: /homefolder/CLC_Data

You can easily add more locations to the **Navigation Area**:

File | **New** | **Location (** **)**

This will bring up a dialog where you can navigate to the folder you wish to use as your new location (see figure 3.4).

When you click **Open**, the new location is added to the **Navigation Area** as shown in figure 3.5.

The name of the new location will be the name of the folder selected for the location. To see where the folder is located on your computer, place your mouse cursor on the location icon ( ) for second. This will show the path to the location.

Sharing data is possible of you add a location on a network drive. The procedure is similar to the one described above. When you add a location on a network drive or a removable drive, the

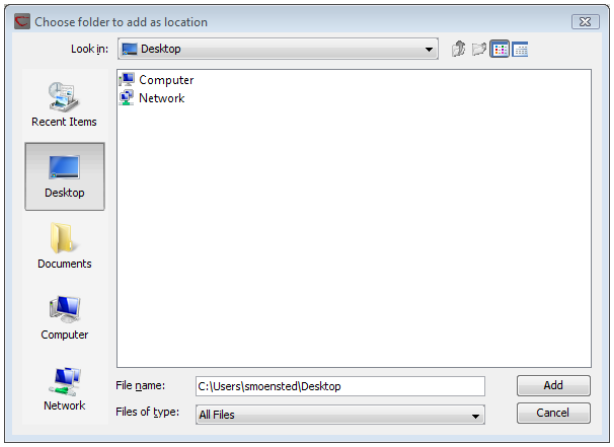


Figure 3.4: *Navigating to a folder to use as a new location.*

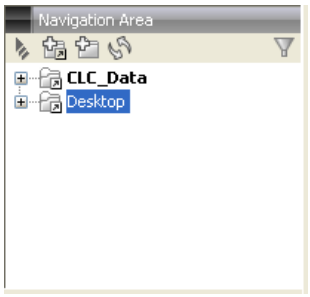


Figure 3.5: *The new location has been added.*

location will appear *inactive* when you are not connected. Once you connect to the drive again, click **Update All** ( ) and it will become active (note that there will be a few seconds' delay from you connect).

Opening data

The elements in the **Navigation Area** are opened by :

Double-click the element

or **Click the element | Show ( ) in the Toolbar | Select the desired way to view the element**

This will open a view in the **View Area**, which is described in section 3.2.

Adding data

Data can be added to the **Navigation Area** in a number of ways. Files can be imported from the file system (see chapter 7). Furthermore, an element can be added by dragging it into the **Navigation Area**. This could be views that are open, elements on lists, e.g. search hits or sequence lists, and files located on your computer. Finally, you can add data by adding a new location (see section 3.1.1).

If a file or another element is dropped on a folder, it is placed at the bottom of the folder. If it is dropped on another element, it will be placed just below that element.

If the element already exists in the **Navigation Area**, you will be asked whether you wish to create

a copy.

3.1.2 Create new folders

In order to organize your files, they can be placed in folders. Creating a new folder can be done in two ways:

right-click an element in the Navigation Area | **New** | **Folder ( )**

or **File** | **New** | **Folder ( )**

If a folder is selected in the **Navigation Area** when adding a new folder, the new folder is added at the bottom of this folder. If an element is selected, the new folder is added right above that element.

You can move the folder manually by selecting it and dragging it to the desired destination.

3.1.3 Sorting folders

You can sort the elements in a folder alphabetically:

right-click the folder | **Sort Folder**

On Windows, subfolders will be placed at the top of the folder, and the rest of the elements will be listed below in alphabetical order. On Mac, both subfolders and other elements are listed together in alphabetical order.

3.1.4 Multiselecting elements

Multiselecting elements means that you select more than one element at the same time. This can be done in the following ways:

- Holding down the <Ctrl> key (⌘ on Mac) while clicking on multiple elements selects the elements that have been clicked.
- Selecting one element, and selecting another element while holding down the <Shift> key selects all the elements listed between the two locations (the two end locations included).
- Selecting one element, and moving the curser with the arrow-keys while holding down the <Shift> key, enables you to increase the number of elements selected.

3.1.5 Moving and copying elements

Elements can be moved and copied in several ways:

- Using **Copy** ( ), **Cut** () and **Paste** ( ) from the **Edit** menu.
- Using Ctrl + C (⌘ + C on Mac), Ctrl + X (⌘ + X on Mac) and Ctrl + V (⌘ + V on Mac).
- Using **Copy** ( ), **Cut** ( ) and **Paste** ( ) in the **Toolbar**.
- Using drag and drop to move elements.

- Using drag and drop while pressing Ctrl / Command to copy elements.

In the following, all of these possibilities for moving and copying elements are described in further detail.

Copy, cut and paste functions

Copies of elements and folders can be made with the copy/paste function which can be applied in a number of ways:

> **select the files to copy** | **right-click one of the selected files** | **Copy ()** | **right-click the location to insert files into** | **Paste ( )**

or **select the files to copy** | **Ctrl + C (⌘ + C on Mac)** | **select where to insert files** | **Ctrl + P (⌘ + P on Mac)**

or **select the files to copy** | **Edit in the Menu Bar** | **Copy ( )** | **select where to insert files** | **Edit in the Menu Bar** | **Paste ( )**

If there is already an element of that name, the pasted element will be renamed by appending a number at the end of the name.

Elements can also be moved instead of copied. This is done with the cut/paste function:

> **select the files to cut** | **right-click one of the selected files** | **Cut ( )** | **right-click the location to insert files into** | **Paste ( )**

or **select the files to cut** | **Ctrl + X (⌘ + X on Mac)** | **select where to insert files** | **Ctrl + V (⌘ + V on Mac)**

When you have cut the element, it is "greyed out" until you activate the paste function. If you change your mind, you can revert the cut command by copying another element.

Note that if you move data between locations, the original data is kept. This means that you are essentially doing a copy instead of a move operation.

Move using drag and drop

Using drag and drop in the **Navigation Area**, as well as in general, is a four-step process:

> **click the element** | **click on the element again, and hold left mouse button** | **drag the element to the desired location** | **let go of mouse button**

This allows you to:

- Move elements between different folders in the **Navigation Area**
- Drag from the **Navigation Area** to the **View Area**: A new view is opened in an existing **View Area** if the element is dragged from the **Navigation Area** and dropped next to the tab(s) in that **View Area**.
- Drag from the **View Area** to the **Navigation Area**: The element, e.g. a sequence, alignment, search report etc. is saved where it is dropped. If the element already exists, you are asked whether you want to save a copy. You drag from the **View Area** by dragging the tab of the desired element.

Use of drag and drop is supported throughout the program, also to open and re-arrange views (see section 3.2.6).

Note that if you move data between locations, the original data is kept. This means that you are essentially doing a copy instead of a move operation.

Copy using drag and drop

To copy instead of move using drag and drop, hold the Ctrl (⌘ on Mac) key while dragging:

> **click the element | click on the element again, and hold left mouse button | drag the element to the desired location | press Ctrl (⌘ on Mac) while you let go of mouse button release the Ctrl/⌘ button**

3.1.6 Change element names

This section describes two ways of changing the names of sequences in the **Navigation Area**. In the first part, the sequences themselves are not changed - it's their representation that changes. The second part describes how to change the name of the element.

Change how sequences are displayed

Sequence elements can be displayed in the **Navigation Area** with different types of information:

- Name (this is the default information to be shown).
- Accession (sequences downloaded from databases like GenBank have an accession number).
- Latin name.
- Latin name (accession).
- Common name.
- Common name (accession).

Whether sequences can be displayed with this information depends on their origin. Sequences that you have created yourself or imported might not include this information, and you will only be able to see them represented by their name. However, sequences downloaded from databases like GenBank will include this information. To change how sequences are displayed:

> **right-click any element or folder in the Navigation Area | Sequence Representation | select format**

This will only affect sequence elements, and the display of other types of elements, e.g. alignments, trees and external files, will be not be changed. If a sequence does not have this information, there will be no text next to the sequence icon.

Rename element

Renaming a folder or an element in the **Navigation Area** can be done in three different ways:

select the element | **Edit in the Menu Bar** | **Rename**

or **select the element** | **F2**

click the element once | **wait one second** | **click the element again**

When you can rename the element, you can see that the text is selected and you can move the cursor back and forth in the text. When the editing of the name has finished; press **Enter** or select another element in the **Navigation Area**. If you want to discard the changes instead, press the **Esc**-key.

For renaming annotations instead of folders or elements, see section 10.3.3.

3.1.7 Delete elements

Deleting a folder or an element can be done in two ways:

right-click the element | **Delete (** **)**

or **select the element** | **press Delete key**

This will cause the element to be moved to the **Recycle Bin** ( ) where it is kept until the recycle bin is emptied. This means that you can recover deleted elements later on.

For deleting annotations instead of folders or elements, see section 10.3.4.

Restore Deleted Elements

The elements in the **Recycle Bin** ( ) can be restored by dragging the elements with the mouse into the folder where they used to be.

If you have deleted large amounts of data taking up very much disk space, you can free this disk space by emptying the **Recycle Bin** ( ):

Edit in the Menu Bar | **Empty Recycle Bin (** **)**

Note! This cannot be undone, and you will therefore not be able to recover the data present in the recycle bin when it was emptied.

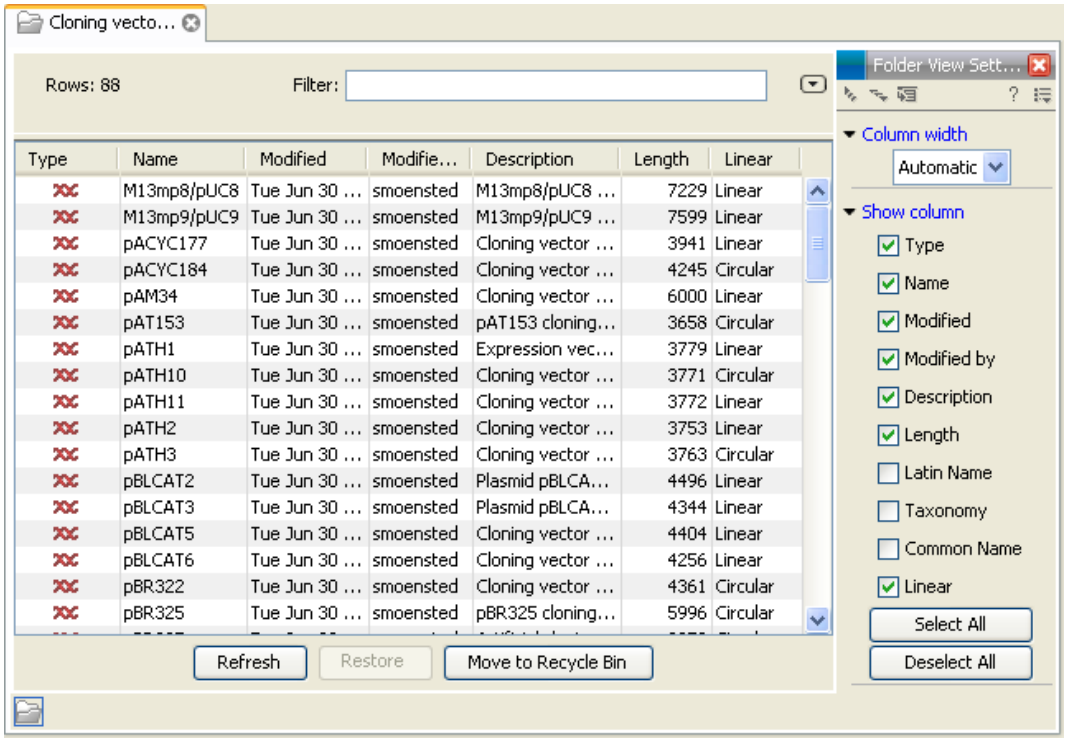
3.1.8 Show folder elements in a table

A location or a folder might contain large amounts of elements. It is possible to view their elements in the **View Area**:

select a folder or location | **Show (** **) in the Toolbar** | **Contents (** **)**

An example is shown in figure 3.6.

When the elements are shown in the view, they can be sorted by clicking the heading of each of the columns. You can further refine the sorting by pressing Ctrl (⌘ on Mac) while clicking the heading of another column.



Type	Name	Modified	Modifie...	Description	Length	Linear
	M13mp8/pUC8	Tue Jun 30 ...	smoensted	M13mp8/pUC8 ...	7229	Linear
	M13mp9/pUC9	Tue Jun 30 ...	smoensted	M13mp9/pUC9 ...	7599	Linear
	pACYC177	Tue Jun 30 ...	smoensted	Cloning vector ...	3941	Linear
	pACYC184	Tue Jun 30 ...	smoensted	Cloning vector ...	4245	Circular
	pAM34	Tue Jun 30 ...	smoensted	Cloning vector ...	6000	Linear
	pAT153	Tue Jun 30 ...	smoensted	pAT153 cloning...	3658	Circular
	pATH1	Tue Jun 30 ...	smoensted	Expression vec...	3779	Linear
	pATH10	Tue Jun 30 ...	smoensted	Cloning vector ...	3771	Circular
	pATH11	Tue Jun 30 ...	smoensted	Cloning vector ...	3772	Linear
	pATH2	Tue Jun 30 ...	smoensted	Cloning vector ...	3753	Linear
	pATH3	Tue Jun 30 ...	smoensted	Cloning vector ...	3763	Circular
	pBLCAT2	Tue Jun 30 ...	smoensted	Plasmid pBLCA...	4496	Linear
	pBLCAT3	Tue Jun 30 ...	smoensted	Plasmid pBLCA...	4344	Linear
	pBLCAT5	Tue Jun 30 ...	smoensted	Cloning vector ...	4404	Linear
	pBLCAT6	Tue Jun 30 ...	smoensted	Cloning vector ...	4256	Linear
	pBR322	Tue Jun 30 ...	smoensted	Cloning vector ...	4361	Circular
	pBR325	Tue Jun 30 ...	smoensted	pBR325 cloning...	5996	Circular

Figure 3.6: *Viewing the elements in a folder.*

Sorting the elements in a view does not affect the ordering of the elements in the **Navigation Area**.

Note! The view only displays one "layer" at a time: the content of subfolders is not visible in this view. Also note that only sequences have the full span of information like organism etc.

Batch edit folder elements

You can select a number of elements in the table, right-click and choose **Edit** to batch edit the elements. In this way, you can change the e.g. the description or common name of several elements in one go.

In figure 3.7 you can see an example where the common name of five sequence are renamed in one go. In this example, a dialog with a text field will be shown, letting you enter a new common name for these five sequences. **Note!** This information is directly saved and you cannot undo.

3.2 View Area

The **View Area** is the right-hand part of the screen, displaying your current work. The **View Area** may consist of one or more **Views**, represented by tabs at the top of the **View Area**.

This is illustrated in figure 3.8.

The tab concept is central to working with *CLC Protein Workbench*, because several operations can be performed by dragging the tab of a view, and extended right-click menus can be activated from the tabs.

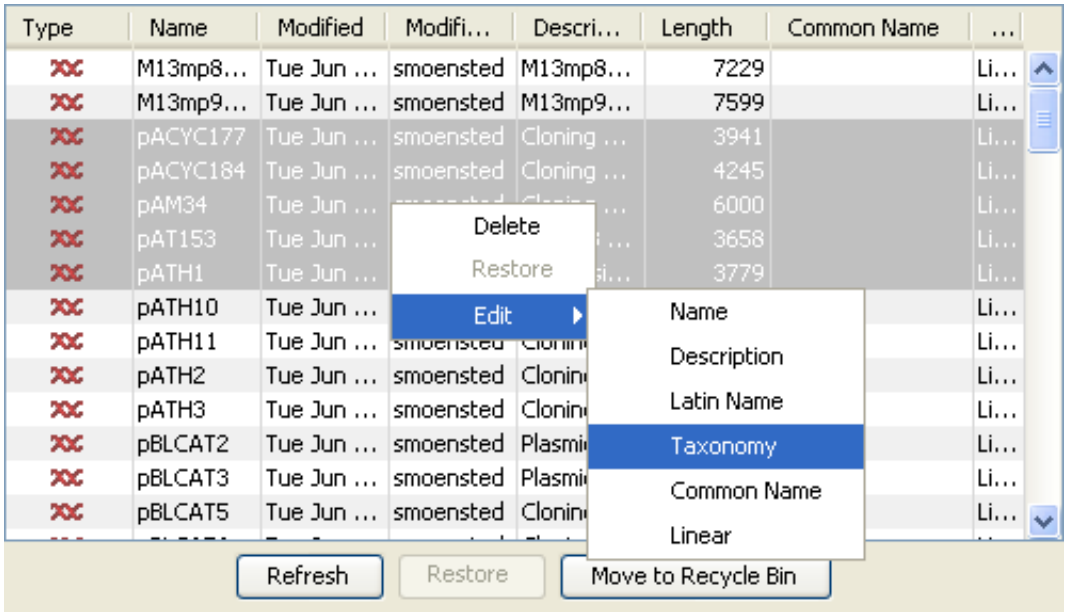


Figure 3.7: *Changing the common name of five sequences.*

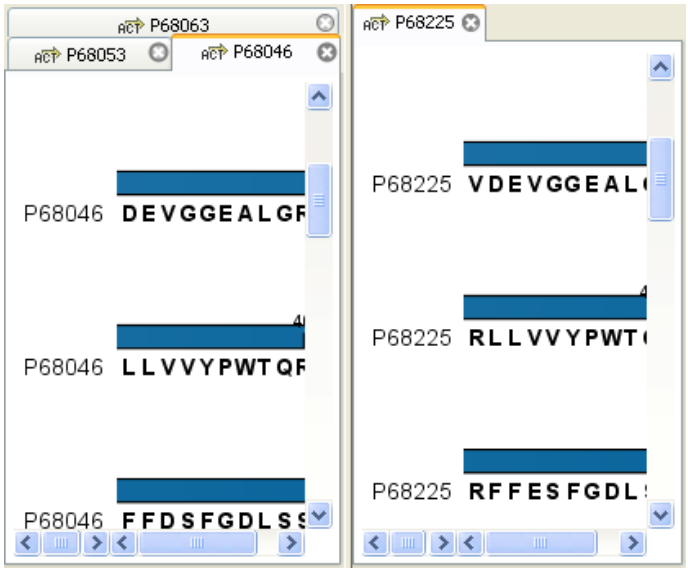


Figure 3.8: *A View Area can enclose several views, each view is indicated with a tab (see right view, which shows protein P68225). Furthermore, several views can be shown at the same time (in this example, four views are displayed).*

This chapter deals with the handling of views inside a **View Area**. Furthermore, it deals with rearranging the views.

Section 3.3 deals with the zooming and selecting functions.

3.2.1 Open view

Opening a view can be done in a number of ways:

double-click an element in the Navigation Area

or **select an element in the Navigation Area | File | Show | Select the desired way to view the element**

or **select an element in the Navigation Area | Ctrl + O (⌘ + B on Mac)**

Opening a view while another view is already open, will show the new view in front of the other view. The view that was already open can be brought to front by clicking its tab.

Note! If you right-click an open tab of any element, click **Show**, and then choose a different view of the same element, this new view is automatically opened in a split-view, allowing you to see both views.

See section 3.1.5 for instructions on how to open a view using drag and drop.

3.2.2 Show element in another view

Each element can be shown in different ways. A sequence, for example, can be shown as linear, circular, text etc.

In the following example, you want to see a sequence in a circular view. If the sequence is already open in a view, you can change the view to a circular view:

Click Show As Circular ( ) at the lower left part of the view

The buttons used for switching views are shown in figure 3.9).

Figure 3.9: *The buttons shown at the bottom of a view of a nucleotide sequence. You can click the buttons to change the view to e.g. a circular view or a history view.*

If the sequence is already open in a linear view ( ), and you wish to see both a circular and a linear view, you can split the views very easily:

Press Ctrl (⌘ on Mac) while you | Click Show As Circular ( ) at the lower left part of the view

This will open a split view with a linear view at the bottom and a circular view at the top (see 10.5).

You can also show a circular view of a sequence without opening the sequence first:

Select the sequence in the Navigation Area | Show ( ) | As Circular ( )

3.2.3 Close views

When a view is closed, the **View Area** remains open as long as there is at least one open view.

A view is closed by:

right-click the tab of the View | Close

or **select the view | Ctrl + W**

or **hold down the Ctrl-button | Click the tab of the view while the button is pressed**

By right-clicking a tab, the following close options exist. See figure 3.10

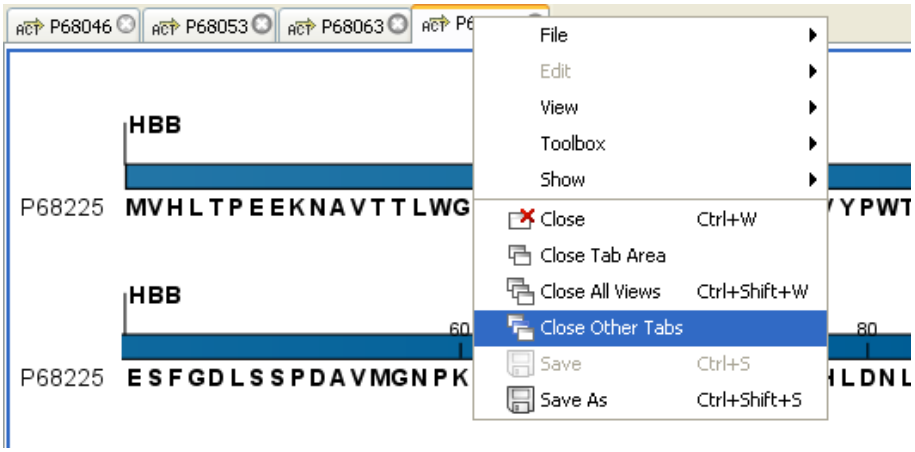


Figure 3.10: *By right-clicking a tab, several close options are available.*

- **Close.** See above.
- **Close Tab Area.** Closes all tabs in the tab area.
- **Close All Views.** Closes all tabs, in all tab areas. Leaves an empty workspace.
- **Close Other Tabs.** Closes all other tabs, in all tab areas, except the one that is selected.

3.2.4 Save changes in a view

When changes are made in a view, the text on the tab appears *bold and italic* (on Mac it is indicated by an * before the name of the tab). This indicates that the changes are not saved. The **Save** function may be activated in two ways:

Click the tab of the view you want to save | **Save (**) **in the toolbar.**

or **Click the tab of the view you want to save** | **Ctrl + S (⌘ + S on Mac)**

If you close a view containing an element that has been changed since you opened it, you are asked if you want to save.

When saving a new view that has not been opened from the Navigation Area (e.g. when opening a sequence from a list of search hits), a save dialog appears (figure 3.11).

In the dialog you select the folder in which you want to save the element.

After naming the element, press **OK**

3.2.5 Undo/Redo

If you make a change in a view, e.g. remove an annotation in a sequence or modify a tree, you can undo the action. In general, **Undo** applies to all changes you can make when right-clicking in a view. **Undo** is done by:

Click undo ( ) **in the Toolbar**

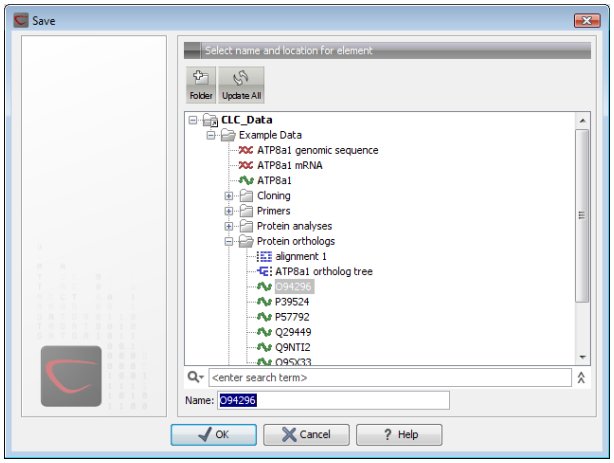


Figure 3.11: *Save dialog.*

or **Edit** | **Undo (** **)**

or **Ctrl + Z**

If you want to undo several actions, just repeat the steps above. To reverse the undo action:

Click the redo icon in the Toolbar

or **Edit** | **Redo (** **)**

or **Ctrl + Y**

Note! Actions in the **Navigation Area**, e.g. renaming and moving elements, cannot be undone. However, you can restore deleted elements (see section 3.1.7).

You can set the number of possible undo actions in the Preferences dialog (see section 5).

3.2.6 Arrange views in View Area

Views are arranged in the **View Area** by their tabs. The order of the **views** can be changed using drag and drop. E.g. drag the tab of one view onto the tab of a another. The tab of the first view is now placed at the right side of the other tab.

If a tab is dragged into a view, an area of the view is made gray (see fig. 3.12) illustrating that the view will be placed in this part of the **View Area**.

The results of this action is illustrated in figure 3.13.

You can also split a **View Area** horizontally or vertically using the menus.

Splitting horisontally may be done this way:

right-click a tab of the view | **View** | **Split Horizontally (** **)**

This action opens the chosen view below the existing view. (See figure 3.14). When the split is made vertically, the new view opens to the right of the existing view.

Splitting the **View Area** can be undone by dragging e.g. the tab of the bottom view to the tab of the top view. This is marked by a gray area on the top of the view.

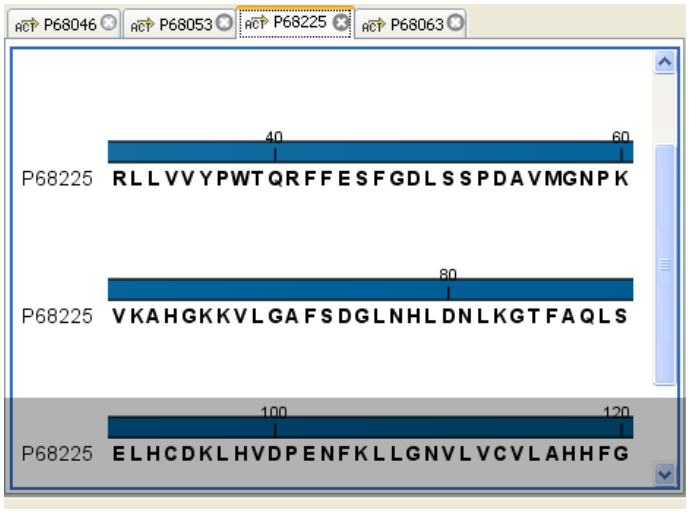


Figure 3.12: *When dragging a view, a gray area indicates where the view will be shown.*

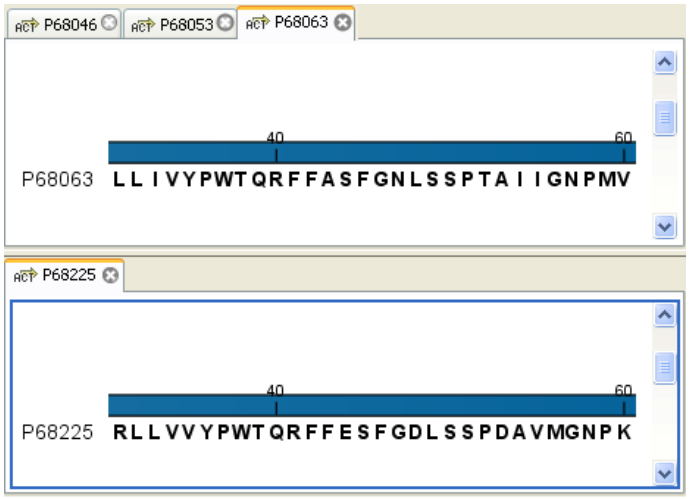


Figure 3.13: *A horizontal split-screen. The two views split the View Area.*

Maximize/Restore size of view

The **Maximize/Restore View** function allows you to see a view in maximized mode, meaning a mode where no other **views** nor the **Navigation Area** is shown.

Maximizing a view can be done in the following ways:

select view | **Ctrl + M**

or **select view** | **View** | **Maximize/restore View (▭)**

or **select view** | **right-click the tab** | **View** | **Maximize/restore View (▭)**

or **double-click the tab of view**

The following restores the size of the view:

Ctrl + M

or **View** | **Maximize/restore View (▭)**

or **double-click title of view**

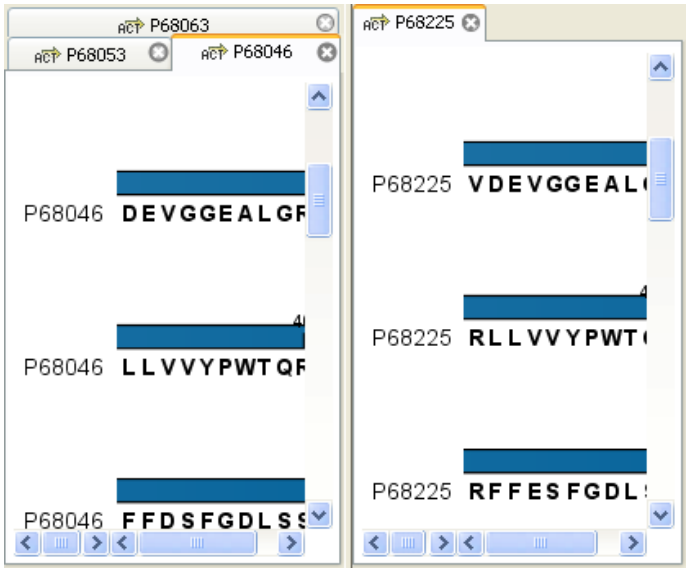


Figure 3.14: *A vertical split-screen.*

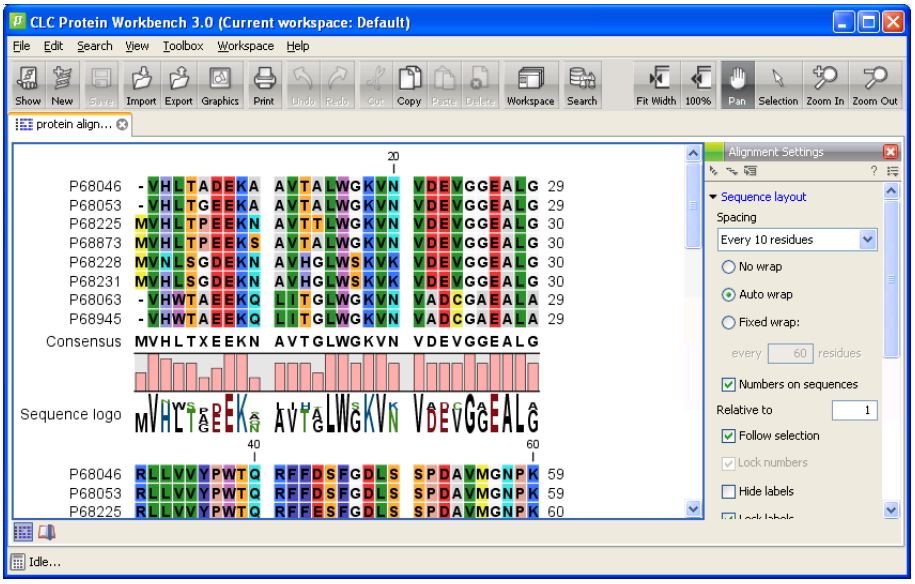


Figure 3.15: *A maximized view. The function hides the Navigation Area and the Toolbox.*

3.2.7 Side Panel

The **Side Panel** allows you to change the way the contents of a view are displayed. The options in the **Side Panel** depend on the kind of data in the view, and they are described in the relevant sections about sequences, alignments, trees etc.

Side Panel are activated in this way:

select the view | **Ctrl + U (⌘ + U on Mac)**

or **right-click the tab of the view** | **View** | **Show/Hide Side Panel (** **)**

Note! Changes made to the **Side Panel** will not be saved when you save the view. See how to save the changes in the **Side Panel** in chapter 5 .

The **Side Panel** consists of a number of groups of preferences (depending on the kind of data

being viewed), which can be expanded and collapsed by clicking the header of the group. You can also expand or collapse all the groups by clicking the icons ( )/ ( ) at the top.

3.3 Zoom and selection in View Area

The mode toolbar items in the right side of the **Toolbar** apply to the function of the mouse pointer. When e.g. **Zoom Out** is selected, you zoom out each time you click in a view where zooming is relevant (texts, tables and lists cannot be zoomed). The chosen mode is active until another mode toolbar item is selected. (**Fit Width** and **Zoom to 100%** do not apply to the mouse pointer.)

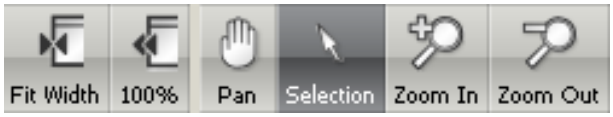


Figure 3.16: *The mode toolbar items.*

3.3.1 Zoom In

There are four ways of **Zooming In**:

Click Zoom In ( ) in the toolbar | **click the location in the view that you want to zoom in on**.

or **Click Zoom In ( ) in the toolbar** | **click-and-drag a box around a part of the view** | **the view now zooms in on the part you selected**

or **Press '+' on your keyboard**

The last option for zooming in is only available if you have a mouse with a scroll wheel:

or **Press and hold Ctrl (⌘ on Mac)** | **Move the scroll wheel on your mouse forward**

When you choose the Zoom In mode, the mouse pointer changes to a magnifying glass to reflect the mouse mode.

Note! You might have to click in the view before you can use the keyboard or the scroll wheel to zoom.

If you press the **Shift** button on your keyboard while clicking in a **View**, the zoom function is reversed. Hence, clicking on a sequence in this way while the **Zoom In** mode toolbar item is selected, zooms out instead of zooming in.

3.3.2 Zoom Out

It is possible to zoom out, step by step, on a sequence:

Click Zoom Out ( ) in the toolbar | **click in the view until you reach a satisfying zoomlevel**.

or **Press '-' on your keyboard**

The last option for zooming out is only available if you have a mouse with a scroll wheel:

or **Press and hold Ctrl (⌘ on Mac)** | **Move the scroll wheel on your mouse backwards**

When you choose the Zoom Out mode, the mouse pointer changes to a magnifying glass to reflect the mouse mode.

Note! You might have to click in the view before you can use the keyboard or the scroll wheel to zoom.

If you want to get a quick overview of a sequence or a tree, use the **Fit Width** function instead of the **Zoom Out** function.

If you press **Shift** while clicking in a **View**, the zoom function is reversed. Hence, clicking on a sequence in this way while the **Zoom Out** mode toolbar item is selected, zooms in instead of zooming out.

3.3.3 Fit Width

The **Fit Width** ( ) function adjusts the content of the **View** so that both ends of the sequence, alignment, or tree is visible in the **View** in question. (This function does not change the mode of the mouse pointer.)

3.3.4 Zoom to 100%

The **Zoom to 100%** ( ) function zooms the content of the **View** so that it is displayed with the highest degree of detail. (This function does not change the mode of the mouse pointer.)

3.3.5 Move

The Move mode allows you to drag the content of a **View**. E.g. if you are studying a sequence, you can click anywhere in the sequence and hold the mouse button. By moving the mouse you move the sequence in the **View**.

3.3.6 Selection

The Selection mode ( ) is used for selecting in a **View** (selecting a part of a sequence, selecting nodes in a tree etc.). It is also used for moving e.g. branches in a tree or sequences in an alignment.

When you make a selection on a sequence or in an alignment, the location is shown in the bottom right corner of the screen. E.g. ’23^24’ means that the selection is between two residues. ’23’ means that the residue at position 23 is selected, and finally ’23..25’ means that 23, 24 and 25 are selected. By holding ctrl / ⌘ you can make multiple selections.

3.3.7 Changing compactness

There is a shortcut way of changing the compactness setting for read mappings:

or **Press and hold Alt key** | **Scroll using your mouse wheel or touchpad**

3.4 Toolbox and Status Bar

The **Toolbox** is placed in the left side of the user interface of *CLC Protein Workbench* below the **Navigation Area**.

The **Toolbox** shows a **Processes tab** and a **Toolbox tab**.

3.4.1 Processes

By clicking the **Processes** tab, the **Toolbox** displays previous and running processes, e.g. an NCBI search or a calculation of an alignment. The running processes can be stopped, paused, and resumed by clicking the small icon ( ) next to the process (see figure 3.17).

Running and paused processes are not deleted.

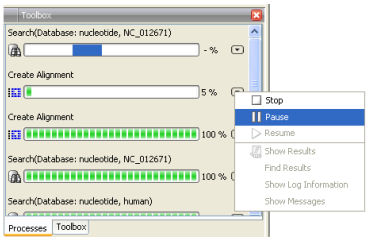


Figure 3.17: *A database search and an alignment calculation are running. Clicking the small icon next to the process allow you to stop, pause and resume processes.*

Besides the options to stop, pause and resume processes, there are some extra options for *a selected number* of the tools running from the Toolbox:

- **Show results**. If you have chosen to save the results (see section 9.1), you will be able to open the results directly from the process by clicking this option.
- **Find results**. If you have chosen to save the results (see section 9.1), you will be able to high-light the results in the Navigation Area.
- **Show Log Information**. This will display a log file showing progress of the process. The log file can also be shown by clicking **Show Log** in the "handle results" dialog where you choose between saving and opening the results.
- **Show Messages**. Some analyses will give you a message when processing your data. The messages are the black dialogs shown in the lower left corner of the Workbench that disappear after a few seconds. You can reiterate the messages that have been shown by clicking this option.

The terminated processes can be removed by:

View | **Remove Terminated Processes ( )**

If you close the program while there are running processes, a dialog will ask if you are sure that you want to close the program. Closing the program will stop the process, and it cannot be restarted when you open the program again.

3.4.2 Toolbox

The content of the **Toolbox** tab in the **Toolbox** corresponds to **Toolbox** in the **Menu Bar**.

The **Toolbox** can be hidden, so that the **Navigation Area** is enlarged and thereby displays more elements:

View | **Show/Hide Toolbox**

The tools in the toolbox can be accessed by double-clicking or by dragging elements from the **Navigation Area** to an item in the **Toolbox**.

3.4.3 Status Bar

As can be seen from figure 3.1, the **Status Bar** is located at the bottom of the window. In the left side of the bar is an indication of whether the computer is making calculations or whether it is idle. The right side of the **Status Bar** indicates the range of the selection of a sequence. (See chapter 3.3.6 for more about the Selection mode button.)

3.5 Workspace

If you are working on a project and have arranged the views for this project, you can save this arrangement using **Workspaces**. A Workspace remembers the way you have arranged the views, and you can switch between different workspaces.

The **Navigation Area** always contains the same data across **Workspaces**. It is, however, possible to open different folders in the different **Workspaces**. Consequently, the program allows you to display different clusters of the data in separate **Workspaces**.

All **Workspaces** are automatically saved when closing down *CLC Protein Workbench*. The next time you run the program, the **Workspaces** are reopened exactly as you left them.

Note! It is not possible to run more than one version of *CLC Protein Workbench* at a time. Use two or more **Workspaces** instead.

3.5.1 Create Workspace

When working with large amounts of data, it might be a good idea to split the work into two or more **Workspaces**. As default the *CLC Protein Workbench* opens one **Workspace**. Additional **Workspaces** are created in the following way:

Workspace in the Menu Bar) | **Create Workspace** | **enter name of Workspace** | **OK**

When the new **Workspace** is created, the heading of the program frame displays the name of the new **Workspace**. Initially, the selected elements in the **Navigation Area** is collapsed and the **View Area** is empty and ready to work with. (See figure 3.18).

3.5.2 Select Workspace

When there is more than one **Workspace** in the *CLC Protein Workbench*, there are two ways to switch between them:

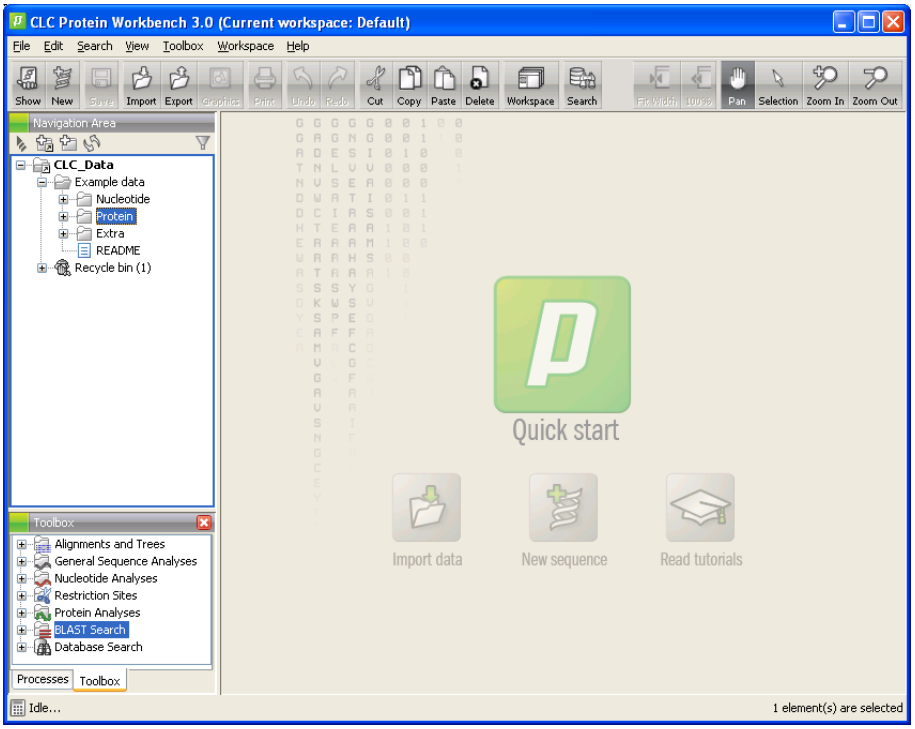


Figure 3.18: *An empty Workspace.*

Workspace ( ) in the Toolbar | Select the Workspace to activate

or **Workspace in the Menu Bar | Select Workspace ( ) | choose which Workspace to activate | OK**

The name of the selected **Workspace** is shown after "*CLC Protein Workbench*" at the top left corner of the main window, in figure 3.18 it says: (default).

3.5.3 Delete Workspace

Deleting a **Workspace** can be done in the following way:

Workspace in the Menu Bar | Delete Workspace | choose which Workspace to delete | OK

Note! Be careful to select the right **Workspace** when deleting. The delete action cannot be undone. (However, no data is lost, because a workspace is only a representation of data.)

It is not possible to delete the default workspace.

3.6 List of shortcuts

The keyboard shortcuts in *CLC Protein Workbench* are listed below.

Action	Windows/Linux	Mac OS X
Adjust selection	Shift + arrow keys	Shift + arrow keys
Change between tabs[1]	Ctrl + tab	Ctrl + Page Up/Down
Close	Ctrl + W	⌘ + W
Close all views	Ctrl + Shift + W	⌘ + Shift + W
Copy	Ctrl + C	⌘ + C
Cut	Ctrl + X	⌘ + X
Delete	Delete	Delete or ⌘ + Backspace
Exit	Alt + F4	⌘ + Q
Export	Ctrl + E	⌘ + E
Export graphics	Ctrl + G	⌘ + G
Find Next Conflict	Space or .	Space or .
Find Previous Conflict	,	,
Help	F1	F1
Import	Ctrl + I	⌘ + I
Maximize/restore size of View	Ctrl + M	⌘ + M
Move gaps in alignment	Ctrl + arrow keys	⌘ + arrow keys
Navigate sequence views	arrow keys	arrow keys
New Folder	Ctrl + Shift + N	⌘ + Shift + N
New Sequence	Ctrl + N	⌘ + N
View	Ctrl + O	⌘ + O
Paste	Ctrl + V	⌘ + V
Print	Ctrl + P	⌘ + P
Redo	Ctrl + Y	⌘ + Y
Rename	F2	F2
Save	Ctrl + S	⌘ + S
Search local data	Ctrl + F	⌘ + F
Search within a sequence	Ctrl + Shift + F	⌘ + Shift + F
Search NCBI	Ctrl + B	⌘ + B
Search UniProt	Ctrl + Shift + U	⌘ + Shift + U
Select All	Ctrl + A	⌘ + A
Selection Mode	Ctrl + 2	⌘ + 2
Show/hide Side Panel	Ctrl + U	⌘ + U
Sort folder	Ctrl + Shift + R	⌘ + Shift + R
Split Horizontally	Ctrl + T	⌘ + T
Split Vertically	Ctrl + J	⌘ + J
Undo	Ctrl + Z	⌘ + Z
User Preferences	Ctrl + K	⌘ + ;
Zoom In Mode	Ctrl + + (plus)	⌘ + 3
Zoom In (without clicking)	+ (plus)	+ (plus)
Zoom Out Mode	Ctrl + - (minus)	⌘ + 4
Zoom Out (without clicking)	- (minus)	- (minus)
Inverse zoom mode	press and hold Shift	press and hold Shift

Combinations of keys and mouse movements are listed below.

[1] On Linux changing tabs is accomplished using Ctrl + Page Up/Page Down

Action	Windows/Linux	Mac OS X	Mouse movement
Maximize View			Double-click the tab of the View
Restore View			Double-click the View title
Reverse zoom function	Shift	Shift	Click in view
Select multiple elements	Ctrl	⌘	Click elements
Select multiple elements	Shift	Shift	Click elements

"Elements" in this context refers to elements and folders in the **Navigation Area** selections on sequences, and rows in tables.

Chapter 4

Searching your data

Contents

4.1 What kind of information can be searched? 83
4.2 Quick search 84
4.2.1 Quick search results 84
4.2.2 Special search expressions 85
4.2.3 Quick search history 86
4.3 Advanced search 86
4.4 Search index 88

There are two ways of doing text-based searches of your data, as described in this chapter:

- **Quick-search** directly from the search field in the **Navigation Area**.
- **Advanced search** which makes it easy to make more specific searches.

In most cases, quick-search will find what you need, but if you need to be more specific in your search criteria, the advanced search is preferable.

4.1 What kind of information can be searched?

Below is a list of the different kinds of information that you can search for (applies to both quick-search and the advanced search).

- **Name.** The name of a sequence, an alignment or any other kind of element. The name is what is displayed in the **Navigation Area** per default.
- **Length.** The length of the sequence.
- **Organism.** Sequences which contain information about organism can be searched. In this way, you could search for e.g. *Homo sapiens* sequences.
- **Database fields.** If your data is stored in a CLC Bioinformatics Database, you will be able to search for custom defined information. Read more in the database user manual.

Only the first item in the list, **Name**, is available for all kinds of data. The rest is only relevant for sequences.

If you wish to perform a search for sequence similarity, use Local BLAST (see section 12.1.3) instead.

4.2 Quick search

At the bottom of the **Navigation Area** there is a text field as shown in figure 4.1).

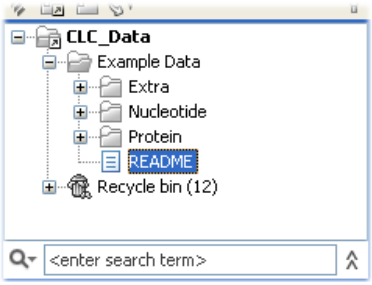


Figure 4.1: *Search simply by typing in the text field and press Enter.*

To search, simply enter a text to search for and press **Enter**.

4.2.1 Quick search results

To show the results, the search pane is expanded as shown in figure 4.2).

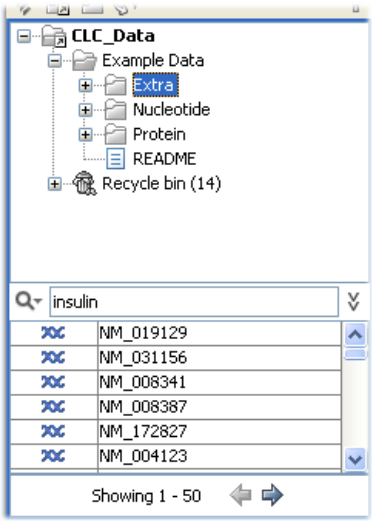


Figure 4.2: *Search results.*

If there are many hits, only the 50 first hits are immediately shown. At the bottom of the pane you can click **Next** () to see the next 50 hits (see figure 4.3).

If a search gives no hits, you will be asked if you wish to search for matches that start with your search term. If you accept this, an asterisk (*) will be appended to the search term.

Pressing the Alt key while you click a search result will high-light the search hit in its folder in the **Navigation Area**.

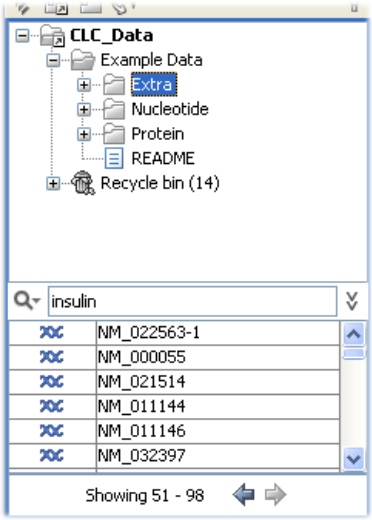


Figure 4.3: *Page two of the search results.*

In the preferences (see 5), you can specify the number of hits to be shown.

4.2.2 Special search expressions

When you write a search term in the search field, you can get help to write a more advanced search expression by pressing **Shift+F1**. This will reveal a list of guides as shown in figure 4.4.

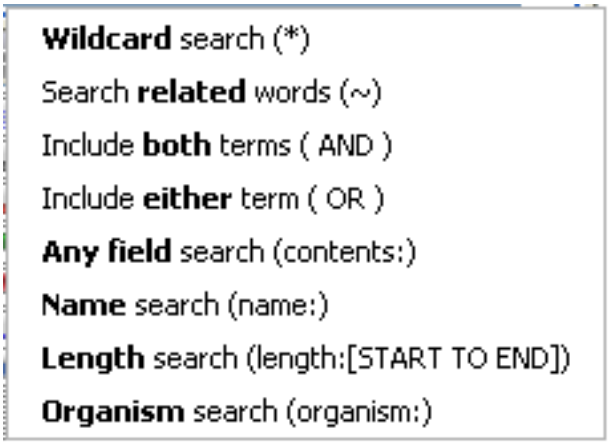

Wildcard search (*)
Search related words (~)
Include both terms (AND)
Include either term (OR)
Any field search (contents:)
Name search (name:)
Length search (length:[START TO END])
Organism search (organism:)

Figure 4.4: *Guides to help create advanced search expressions.*

You can select any of the guides (using mouse or keyboard arrows), and start typing. If you e.g. wish to search for sequences named BRCA1, select "Name search (name:)", and type "BRCA1". Your search expression will now look like this: "name:BRCA1".

The guides available are these:

- **Wildcard search (*)**. Appending an asterisk * to the search term will find matches starting with the term. E.g. searching for "brca*" will find both *brca1* and *brca2*.
- **Search related words ( )**. If you don't know the exact spelling of a word, you can append a question mark to the search term. E.g. "brac1*" will find sequences with a *brca1* gene.

- **Include both terms (AND)**. If you write two search terms, you can define if your results have to match both search terms by combining them with AND. E.g. search for "brca1 AND human" will find sequences where *both* terms are present.
- **Include either term (OR)**. If you write two search terms, you can define that your results have to match either of the search terms by combining them with OR. E.g. search for "brca1 OR brca2" will find sequences where *either* of the terms is present.
- **Name search (name:)**. Search only the name of element.
- **Organism search (organism:)**. For sequences, you can specify the organism to search for. This will look in the "Latin name" field which is seen in the **Sequence Info** view (see section 10.4).
- **Length search (length:[START TO END])**. Search for sequences of a specific length. E.g. search for sequences between 1000 and 2000 residues: "length:1000 TO 2000".

If you do not use this special syntax, you will automatically search for both name, description, organism, etc., and search terms will be combined as if you had put OR between them.

4.2.3 Quick search history

You can access the 10 most recent searches by clicking the icon (🔍) next to the search field (see figure 4.5).

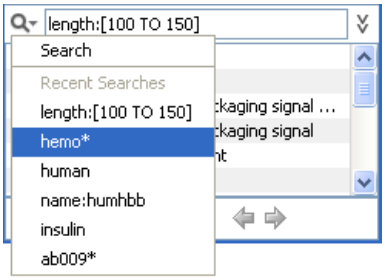


Figure 4.5: *Recent searches.*

Clicking one of the recent searches will conduct the search again.

4.3 Advanced search

As a supplement to the **Quick search** described in the previous section you can use the more advanced search:

Search | Local Search (🔭)

or **Ctrl + F (⌘ + F on Mac)**

This will open the search view as shown in figure 4.6

The first thing you can choose is which location should be searched. All the active locations are shown in this list. You can also choose to search all locations. Read more about locations in section 3.1.1.

Furthermore, you can specify what kind of elements should be searched:

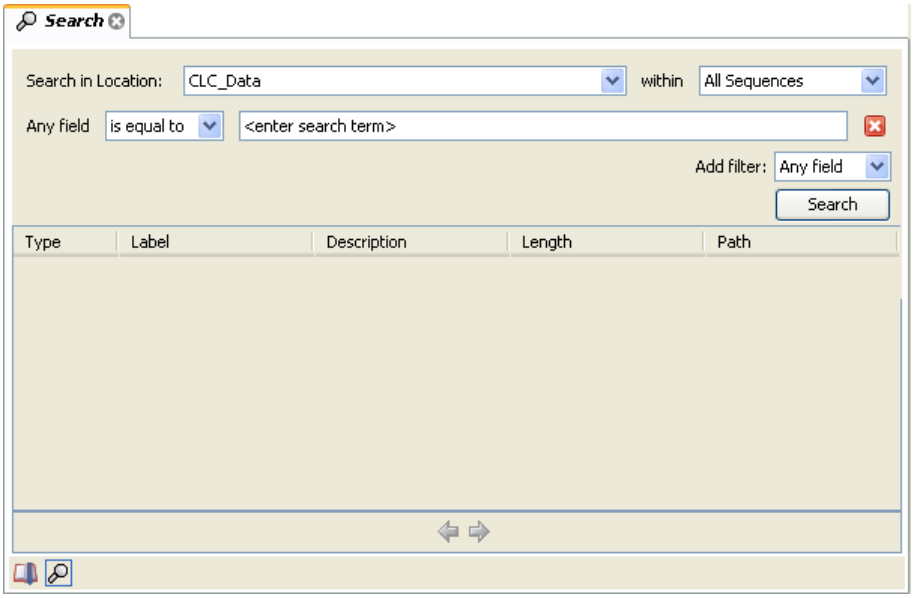


Figure 4.6: *Advanced search.*

- All sequences
- Nucleotide sequences
- Protein sequences
- All data

When searching for sequences, you will also get alignments, sequence lists etc as result, if they contain a sequence which match the search criteria.

Below are the search criteria. First, select a relevant search filter in the **Add filter:** list. For sequences you can search for

- Name
- Length
- Organism

See section 4.2.2 for more information on individual search terms.

For all other data, you can only search for name.

If you use **Any field**, it will search all of the above plus the following:

- Description
- Keywords
- Common name
- Taxonomy name

To see this information for a sequence, switch to the **Element Info** ( ) view (see section 10.4).

For each search line, you can choose if you want the exact term by selecting "is equal to" or if you only enter the start of the term you wish to find (select "begins with").

An example is shown in figure 4.7.

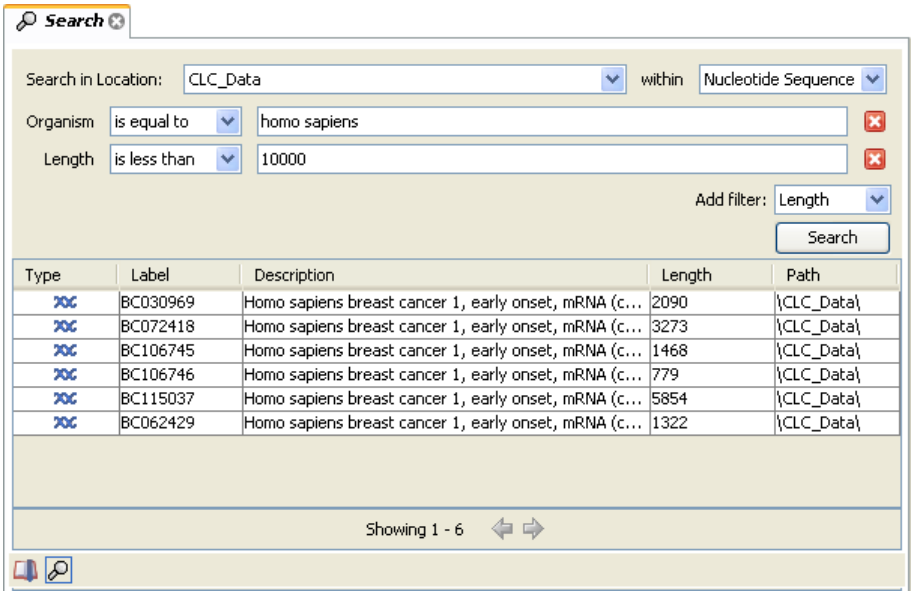


Figure 4.7: *Searching for human sequences shorter than 10,000 nucleotides.*

This example will find human nucleotide sequences (organism is *Homo sapiens*), and it will only find sequences shorter than 10,000 nucleotides.

Note that a search can be saved ( ) for later use. You do not save the search results - only the search parameters. This means that you can easily conduct the same search later on when your data has changed.

4.4 Search index

This section has a technical focus and is not relevant if your search works fine.

However, if you experience problems with your search results: if you do not get the hits you expect, it might be because of an index error.

The *CLC Protein Workbench* automatically maintains an index of all data in all locations in the **Navigation Area**. If this index becomes out of sync with the data, you will experience problems with strange results. In this case, you can rebuild the index:

Right-click the relevant location | **Location** | **Rebuild Index**

This will take a while depending on the size of your data. At any time, the process can be stopped in the process area, see section 3.4.1.

Chapter 5

User preferences and settings

Contents

5.1 General preferences 89
5.2 Default view preferences 90
5.2.1 Number formatting in tables 91
5.2.2 Import and export Side Panel settings 92
5.3 Advanced preferences 93
5.3.1 Default data location 93
5.3.2 NCBI BLAST 94
5.4 Export/import of preferences 94
5.4.1 The different options for export and importing 94
5.5 View settings for the Side Panel 95
5.5.1 Floating Side Panel 97

The first three sections in this chapter deal with the general preferences that can be set for *CLC Protein Workbench* using the **Preferences** dialog. The next section explains how the settings in the **Side Panel** can be saved and applied to other views. Finally, you can learn how to import and export the preferences.

The **Preferences** dialog offers opportunities for changing the default settings for different features of the program.

The **Preferences** dialog is opened in one of the following ways and can be seen in figure 5.1:

Edit | **Preferences (** **)**

or **Ctrl + K (⌘ + ; on Mac)**

5.1 General preferences

The **General** preferences include:

- **Undo Limit.** As default the undo limit is set to 500. By writing a higher number in this field, more actions can be undone. Undo applies to all changes made on sequences, alignments or trees. See section 3.2.5 for more on this topic.

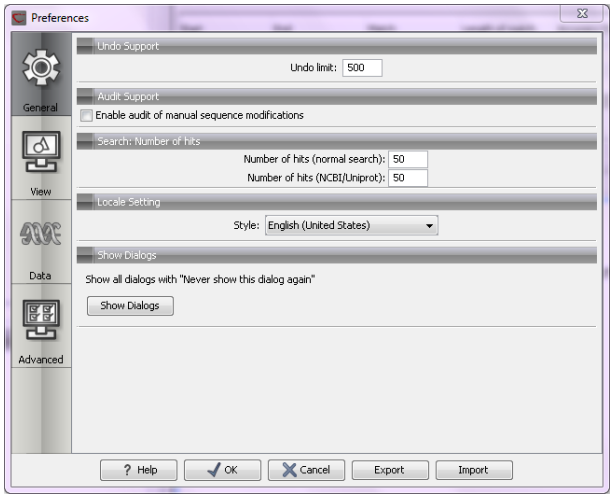


Figure 5.1: *Preferences include General preferences, View preferences, Colors preferences, and Advanced settings.*

- **Audit Support.** If this option is checked, all manual editing of sequences will be marked with an annotation on the sequence (see figure 5.2). Placing the mouse on the annotation will reveal additional details about the change made to the sequence (see figure 5.3). Note that no matter whether **Audit Support** is checked or not, all changes are also recorded in the **History** ( ) (see section 8).
- **Number of hits.** The number of hits shown in *CLC Protein Workbench*, when e.g. searching NCBI. (The sequences shown in the program are not downloaded, until they are opened or dragged/saved into the Navigation Area.
- **Locale Setting.** Specify which country you are located in. This determines how punctation is used in numbers all over the program.
- **Show Dialogs.** A lot of information dialogs have a checkbox: "Never show this dialog again". When you see a dialog and check this box in the dialog, the dialog will not be shown again. If you regret and wish to have the dialog displayed again, click the button in the General Preferences: **Show Dialogs**. Then all the dialogs will be shown again.

Figure 5.2: *Annotations added when the sequence is edited.*

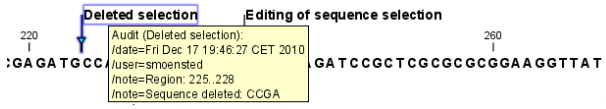


Figure 5.3: *Details of the editing.*

5.2 Default view preferences

There are five groups of default **View** settings:

1. **Toolbar**
2. **Side Panel Location**
3. **New View**
4. **View Format**
5. **User Defined View Settings**.

In general, these are default settings for the user interface.

The **Toolbar preferences** let you choose the size of the toolbar icons, and you can choose whether to display names below the icons.

The **Side Panel Location** setting lets you choose between **Dock in views** and **Float in window**. When docked in view, view preferences will be located in the right side of the view of e.g. an alignment. When floating in window, the side panel can be placed everywhere in your screen, also outside the workspace, e.g. on a different screen. See section 5.5 for more about floating side panels.

The **New view** setting allows you to choose whether the **View preferences** are to be shown automatically when opening a new view. If this option is not chosen, you can press (Ctrl + U (⌘ + U on Mac)) to see the preferences panels of an open view.

The **View Format** allows you to change the way the elements appear in the **Navigation Area**. The following text can be used to describe the element:

- Name (this is the default information to be shown).
- Accession (sequences downloaded from databases like GenBank have an accession number).
- Latin name.
- Latin name (accession).
- Common name.
- Common name (accession).

The **User Defined View Settings** gives you an overview of the different **Side Panel** settings that are saved for each view. See section 5.5 for more about how to create and save style sheets.

If there are other settings beside **CLC Standard Settings**, you can use this overview to choose which of the settings should be used per default when you open a view (see an example in figure 5.4).

In this example, the **CLC Standard Settings** is chosen as default.

5.2.1 Number formatting in tables

In the preferences, you can specify how the numbers should be formatted in tables (see figure 5.5).

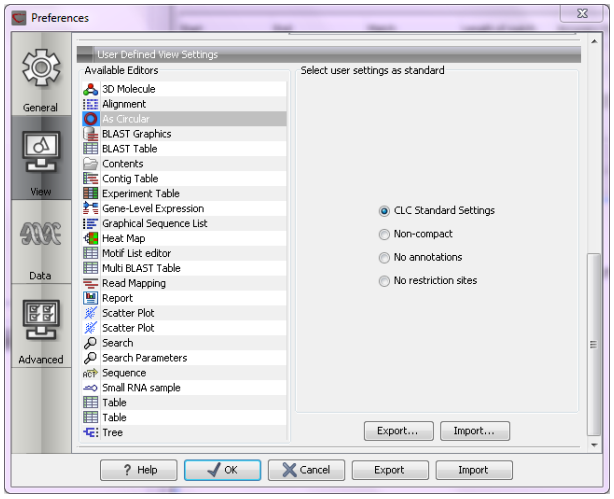


Figure 5.4: *Selecting the default view setting.*

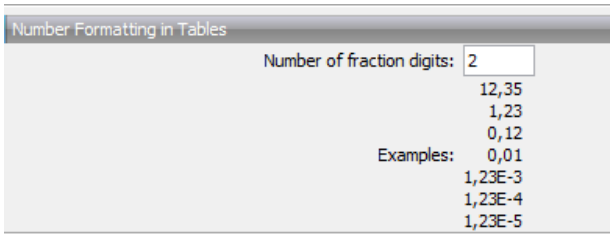


Figure 5.5: *Number formatting of tables.*

The examples below the text field are updated when you change the value so that you can see the effect. After you have changed the preference, you have to re-open your tables to see the effect.

5.2.2 Import and export Side Panel settings

If you have created a special set of settings in the **Side Panel** that you wish to share with other CLC users, you can export the settings in a file. The other user can then import the settings.

To export the **Side Panel** settings, first select the views that you wish to export settings for. Use Ctrl+click (⌘ + click on Mac) or Shift+click to select multiple views. Next click the **Export...**button. Note that there is also another export button at the very bottom of the dialog, but this will export the other settings of the **Preferences** dialog (see section 5.4).

A dialog will be shown (see figure 5.6) that allows you to select which of the settings you wish to export.

When multiple views are selected for export, all the view settings for the views will be shown in the dialog. Click **Export** and you will now be able to define a save folder and name for the exported file. The settings are saved in a file with a .vsf extension (View Settings File).

To import a **Side Panel** settings file, make sure you are at the bottom of the **View** panel of the **Preferences dialog**, and click the **Import...** button. Note that there is also another import button at the very bottom of the dialog, but this will import the other settings of the **Preferences** dialog (see section 5.4).

The dialog asks if you wish to overwrite existing **Side Panel** settings, or if you wish to merge the

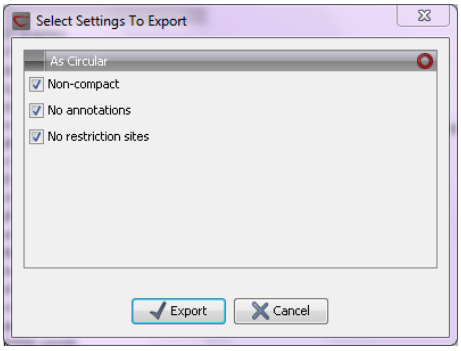


Figure 5.6: *Exporting all settings for circular views.*

imported settings into the existing ones (see figure 5.7).

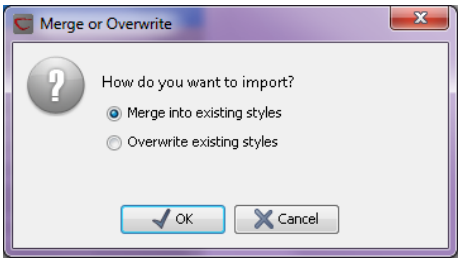


Figure 5.7: *When you import settings, you are asked if you wish to overwrite existing settings or if you wish to merge the new settings into the old ones.*

Note! If you choose to overwrite the existing settings, you will loose all the **Side Panel** settings that you have previously saved.

To avoid confusion of the different import and export options, here is an overview:

- Import and export of **bioinformatics data** such as sequences, alignments etc. (described in section 7.1.1).
- **Graphics** export of the views which creates image files in various formats (described in section 7.3).
- Import and export of **Side Panel Settings** as described above.
- Import and export of all the **Preferences** except the Side Panel settings. This is described in the previous section.

5.3 Advanced preferences

The **Advanced** settings include the possibility to set up a proxy server. This is described in section 1.8.

5.3.1 Default data location

If you have more than one location in the **Navigation Area**, you can choose which location should be the default data location. The default location is used when you e.g. import a file without selecting a folder or element in the **Navigation Area** first. Then the imported element will be placed in the default location.

Note! The default location cannot be removed. You have to select another location as default first.

5.3.2 NCBI BLAST

URL to use for BLAST

It is possible to specify an alternate server URL to use for BLAST searches. The standard URL for the BLAST server at NCBI is: `http://blast.ncbi.nlm.nih.gov/Blast.cgi`.

Note! Be careful to specify a valid URL, otherwise BLAST will not work.

5.4 Export/import of preferences

The user preferences of the *CLC Protein Workbench* can be exported to other users of the program, allowing other users to display data with the same preferences as yours. You can also use the export/import preferences function to backup your preferences.

To export preferences, open the **Preferences** dialog (Ctrl + K (⌘ + ; on Mac)) and do the following:

> **Export** | **Select the relevant preferences** | **Export** | **Choose location for the exported file** | **Enter name of file** | **Save**

Note! The format of exported preferences is .cpf. This notation must be submitted to the name of the exported file in order for the exported file to work.

Before exporting, you are asked about which of the different settings you want to include in the exported file. One of the items in the list is "User Defined View Settings". If you export this, only the information about which of the settings is the default setting for each view is exported. If you wish to export the **Side Panel Settings** themselves, see section 5.2.2.

The process of importing preferences is similar to exporting:

> **Press Ctrl + K (⌘ + ; on Mac) to open Preferences** | **Import** | **Browse to and select the .cpf file** | **Import and apply preferences**

5.4.1 The different options for export and importing

To avoid confusion of the different import and export options, here is an overview:

- Import and export of **bioinformatics data** such as sequences, alignments etc. (described in section 7.1.1).
- **Graphics** export of the views which creates image files in various formats (described in section 7.3).
- Import and export of **Side Panel Settings** as described in the next section.
- Import and export of all the **Preferences** except the Side Panel settings. This is described above.

5.5 View settings for the Side Panel

The **Side Panel** is shown to the right of all views that are opened in *CLC Protein Workbench*. By using the settings in the **Side Panel** you can specify how the layout and contents of the view. Figure 5.8 is an example of the **Side Panel** of a sequence view.

Figure 5.8: *The Side Panel of a sequence contains several groups: Sequence layout, Annotation types, Annotation layout, etc. Several of these groups are present in more views. E.g. Sequence layout is also in the Side Panel of alignment views.*

By clicking the black triangles or the corresponding headings, the groups can be expanded or collapsed. An example is shown in figure 5.9 where the **Sequence layout** is expanded.

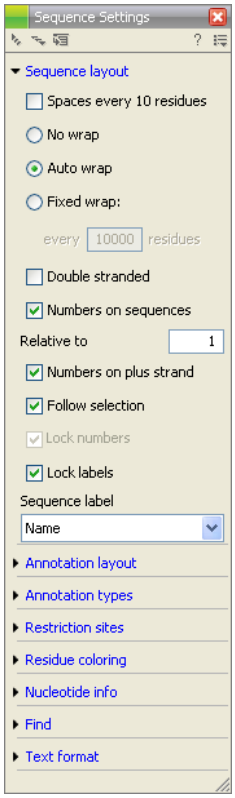


Figure 5.9: *The Sequence layout is expanded.*

The content of the groups is described in the sections where the functionality is explained. E.g. **Sequence Layout** for sequences is described in chapter 10.1.1.

When you have adjusted a view of e.g. a sequence, your settings in the **Side Panel** can be saved. When you open other sequences, which you want to display in a similar way, the saved settings can be applied. The options for saving and applying are available in the top of the **Side Panel** (see figure 5.10).

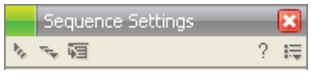


Figure 5.10: *At the top of the Side Panel you can: Expand all groups, Collapse all preferences, Dock/Undock preferences, Help, and Save/Restore preferences.*

To save and apply the saved settings, click ( ) seen in figure 5.10. This opens a menu, where the following options are available:

- **Save Settings.** This brings up a dialog as shown in figure 5.11 where you can enter a name for your settings. Furthermore, by clicking the checkbox **Always apply these settings**, you can choose to use these settings every time you open a new view of this type. If you wish to change which settings should be used per default, open the **Preferences** dialog (see section 5.2).
- **Delete Settings.** Opens a dialog to select which of the saved settings to delete.
- **Apply Saved Settings.** This is a submenu containing the settings that you have previously saved. By clicking one of the settings, they will be applied to the current view. You will also see a number of pre-defined view settings in this submenu. They are meant to be examples of how to use the **Side Panel** and provide quick ways of adjusting the view to common usages. At the bottom of the list of settings you will see **CLC Standard Settings** which represent the way the program was set up, when you first launched it.

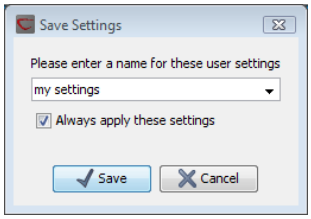


Figure 5.11: *The save settings dialog.*

The settings are specific to the type of view. Hence, when you save settings of a circular view, they will not be available if you open the sequence in a linear view.

If you wish to export the settings that you have saved, this can be done in the **Preferences** dialog under the **View** tab (see section 5.2.2).

The remaining icons of figure 5.10 are used to; **Expand all groups**, **Collapse all groups**, and **Dock/Undock Side Panel**. **Dock/Undock Side Panel** is to make the **Side Panel** "floating" (see below).

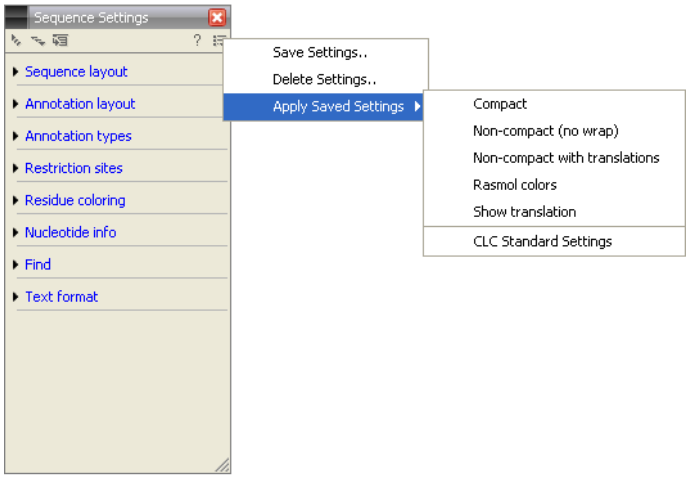


Figure 5.12: *Applying saved settings.*

5.5.1 Floating Side Panel

The Side Panel of the views can be placed in the right side of a view, or it can be floating (see figure 5.13).

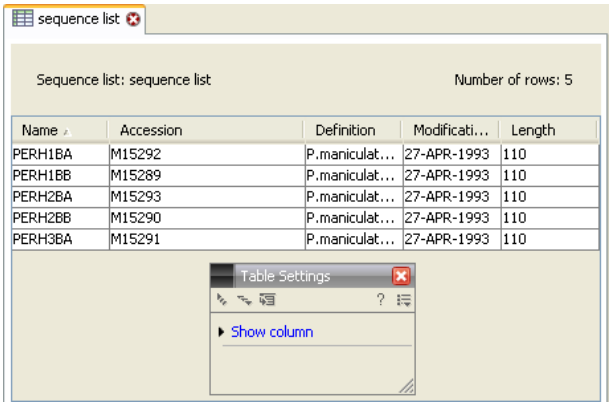


Figure 5.13: *The floating Side Panel can be moved out of the way, e.g. to allow for a wider view of a table.*

By clicking the Dock icon ( ) the floating Side Panel reappear in the right side of the view. The size of the floating Side Panel can be adjusted by dragging the hatched area in the bottom right.

Chapter 6

Printing

Contents

6.1 Selecting which part of the view to print 99
6.2 Page setup 100
6.2.1 Header and footer 101
6.3 Print preview 101

CLC Protein Workbench offers different choices of printing the result of your work.

This chapter deals with printing directly from *CLC Protein Workbench*. Another option for using the graphical output of your work, is to export graphics (see chapter 7.3) in a graphic format, and then import it into a document or a presentation.

All the kinds of data that you can view in the **View Area** can be printed. The *CLC Protein Workbench* uses a WYSIWYG principle: What You See Is What You Get. This means that you should use the options in the Side Panel to change how your data, e.g. a sequence, looks on the screen. When you print it, it will look exactly the same way on print as on the screen.

For some of the views, the layout will be slightly changed in order to be printer-friendly.

It is not possible to print elements directly from the **Navigation Area**. They must first be opened in a view in order to be printed. To print the contents of a view:

select relevant view | Print ( ) in the toolbar

This will show a print dialog (see figure 6.1).

In this dialog, you can:

- Select which part of the view you want to print.
- Adjust **Page Setup**.
- See a print **Preview** window.

These three options are described in the three following sections.

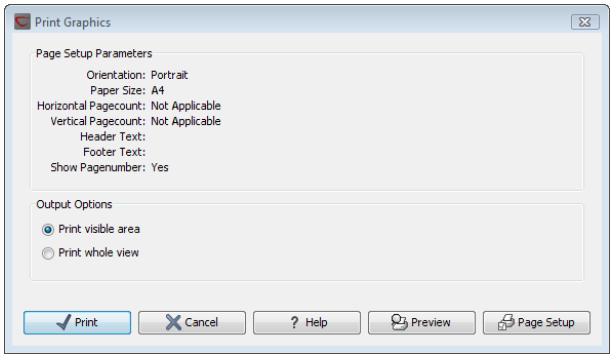


Figure 6.1: *The Print dialog.*

6.1 Selecting which part of the view to print

In the print dialog you can choose to:

- **Print visible area**, or
- **Print whole view**

These options are available for all views that can be zoomed in and out. In figure 6.2 is a view of a circular sequence which is zoomed in so that you can only see a part of it.

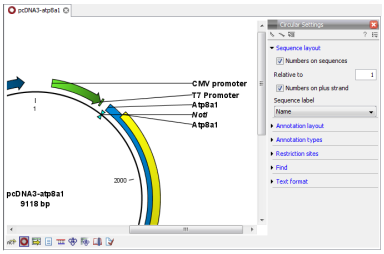


Figure 6.2: *A circular sequence as it looks on the screen.*

When selecting **Print visible area**, your print will reflect the part of the sequence that is *visible* in the view. The result from printing the view from figure 6.2 and choosing **Print visible area** can be seen in figure 6.3.

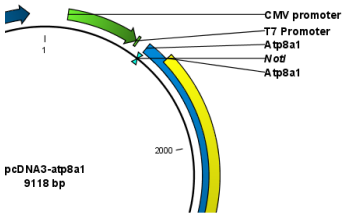


Figure 6.3: *A print of the sequence selecting Print visible area.*

On the other hand, if you select **Print whole view**, you will get a result that looks like figure 6.4. This means that you also print the part of the sequence which is not visible when you have zoomed in.

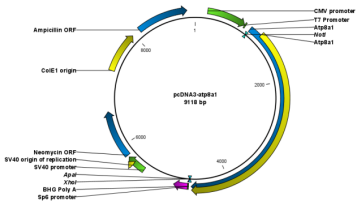


Figure 6.4: *A print of the sequence selecting Print whole view. The whole sequence is shown, even though the view is zoomed in on a part of the sequence.*

6.2 Page setup

No matter whether you have chosen to print the visible area or the whole view, you can adjust page setup of the print. An example of this can be seen in figure 6.5

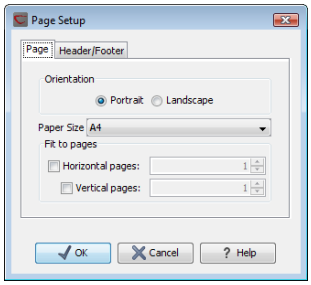


Figure 6.5: *Page Setup.*

In this dialog you can adjust both the setup of the pages and specify a header and a footer by clicking the tab at the top of the dialog.

You can modify the layout of the page using the following options:

- **Orientation**.
  - **Portrait**. Will print with the paper oriented vertically.
  - **Landscape**. Will print with the paper oriented horizontally.
- **Paper size**. Adjust the size to match the paper in your printer.
- **Fit to pages**. Can be used to control how the graphics should be split across pages (see figure 6.6 for an example).
  - **Horizontal pages**. If you set the value to e.g. 2, the printed content will be broken up horizontally and split across 2 pages. This is useful for sequences that are not wrapped
  - **Vertical pages**. If you set the value to e.g. 2, the printed content will be broken up vertically and split across 2 pages.

Note! It is a good idea to consider adjusting view settings (e.g. **Wrap** for sequences), in the **Side Panel** before printing. As explained in the beginning of this chapter, the printed material will look like the view on the screen, and therefore these settings should also be considered when adjusting **Page Setup**.

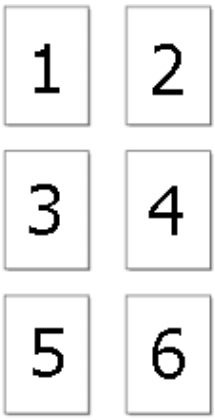


Figure 6.6: *An example where Fit to pages horizontally is set to 2, and Fit to pages vertically is set to 3.*

6.2.1 Header and footer

Click the **Header/Footer** tab to edit the header and footer text. By clicking in the text field for either **Custom header text** or **Custom footer text** you can access the auto formats for header/footer text in **Insert a caret position**. Click either **Date**, **View name**, or **User name** to include the auto format in the header/footer text.

Click **OK** when you have adjusted the **Page Setup**. The settings are saved so that you do not have to adjust them again next time you print. You can also change the **Page Setup** from the **File** menu.

6.3 Print preview

The preview is shown in figure 6.7.

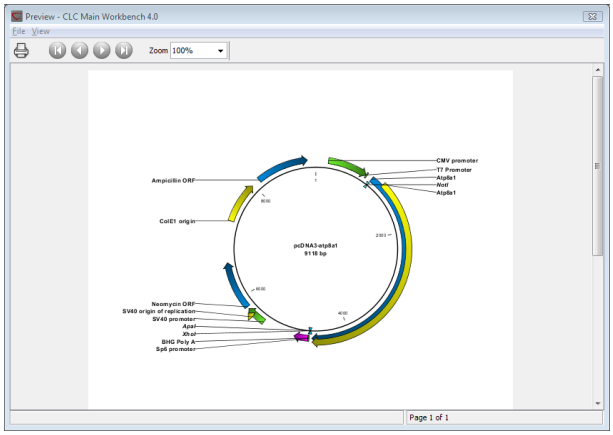


Figure 6.7: *Print preview.*

The **Print preview** window lets you see the layout of the pages that are printed. Use the arrows in the toolbar to navigate between the pages. Click Print ( ) to show the print dialog, which lets you choose e.g. which pages to print.

The **Print preview** window is for preview only - the layout of the pages must be adjusted in the **Page setup**.

Chapter 7

Import/export of data and graphics

Contents

7.1 Bioinformatic data formats . . . 102
7.1.1 Import of bioinformatic data . . . 103
7.1.2 Import Vector NTI data . . . 104
7.1.3 Export of bioinformatics data . . . 107
7.2 External files . . . 109
7.3 Export graphics to files . . . 109
7.3.1 Which part of the view to export . . . 110
7.3.2 Save location and file formats . . . 110
7.3.3 Graphics export parameters . . . 112
7.3.4 Exporting protein reports . . . 113
7.4 Export graph data points to a file . . . 113
7.5 Copy/paste view output . . . 115

CLC Protein Workbench handles a large number of different data formats. All data stored in the Workbench are available in the **Navigation Area**. The data of the **Navigation Area** can be divided into two groups. The data is either one of the different bioinformatic data formats, or it can be an 'external file'. Bioinformatic data formats are those formats which the program can work with, e.g. sequences, alignments and phylogenetic trees. External files are files or links which are stored in *CLC Protein Workbench*, but are opened by other applications, e.g. pdf-files, Microsoft Word files, Open Office spreadsheet files, or links to programs and web-pages etc.

This chapter first deals with importing and exporting data in bioinformatic data formats and as external files. Next comes an explanation of how to export graph data points to a file, and how export graphics.

7.1 Bioinformatic data formats

The different bioinformatic data formats are imported in the same way, therefore, the following description of data import is an example which illustrates the general steps to be followed, regardless of which format you are handling.

7.1.1 Import of bioinformatic data

CLC Protein Workbench has support for a wide range of bioinformatic data such as sequences, alignments etc. See a full list of the data formats in section F.1.

The *CLC Protein Workbench* offers a lot of possibilities to handle bioinformatic data. Read the next sections to get information on how to import different file formats or to import data from a Vector NTI database.

Import using the import dialog

To start the import using the import dialog:

click Import ( ) in the Toolbar

This will show a dialog similar to figure 7.1 (depending on which platform you use). You can change which kind of file types that should be shown by selecting a file format in the **Files of type** box.

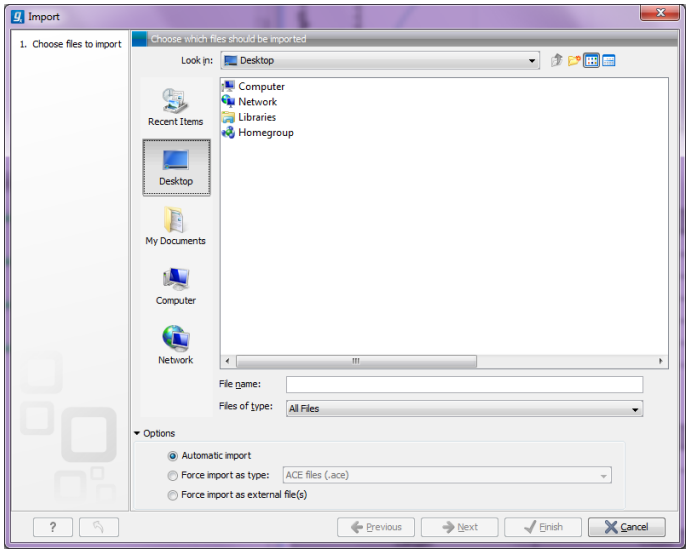


Figure 7.1: *The import dialog.*

Next, select one or more files or folders to import and click **Select**.

This allows you to select a place for saving the result files.

If you import one or more folders, the contents of the folder is automatically imported and placed in that folder in the **Navigation Area**. If the folder contains subfolders, the whole folder structure is imported.

In the import dialog (figure 7.1), there are three import options:

Automatic import This will import the file and *CLC Protein Workbench* will try to determine the format of the file. The format is determined based on the file extension (e.g. SwissProt files have .swp at the end of the file name) in combination with a detection of elements in the file that are specific to the individual file formats. If the file type is not recognized, it will be imported as an external file. In most cases, automatic import will yield a successful result, but if the import goes wrong, the next option can be helpful:

Force import as type This option should be used if *CLC Protein Workbench* cannot successfully determine the file format. By forcing the import as a specific type, the automatic determination of the file format is bypassed, and the file is imported as the type specified.

Force import as external file This option should be used if a file is imported as a bioinformatics file when it should just have been external file. It could be an ordinary text file which is imported as a sequence.

Import using drag and drop

It is also possible to drag a file from e.g. the desktop into the **Navigation Area** of *CLC Protein Workbench*. This is equivalent to importing the file using the **Automatic import** option described above. If the file type is not recognized, it will be imported as an external file.

Import using copy/paste of text

If you have e.g. a text file or a browser displaying a sequence in one of the formats that can be imported by *CLC Protein Workbench*, there is a very easy way to get this sequence into the **Navigation Area**:

Copy the text from the text file or browser | **Select a folder in the Navigation Area** | **Paste ( )**

This will create a new sequence based on the text copied. This operation is equivalent to saving the text in a text file and importing it into the *CLC Protein Workbench*.

If the sequence is not formatted, i.e. if you just have a text like this: "ATGACGAATAGGAGTTC-TAGCTA" you can also paste this into the **Navigation Area**.

Note! Make sure you copy all the relevant text - otherwise *CLC Protein Workbench* might not be able to interpret the text.

7.1.2 Import Vector NTI data

There are several ways of importing your Vector NTI data into the CLC Workbench. The best way to go depends on how your data is currently stored in Vector NTI:

- Your data is stored in the Vector NTI Local Database which can be accessed through Vector NTI Explorer. This is described in the first section below.
- Your data is stored as single files on your computer (just like Word documents etc.). This is described in the second section below.

Import from the Vector NTI Local Database

If your Vector NTI data are stored in a Vector NTI Local Database (as the one shown in figure 7.2), you can import all the data in one step, or you can import selected parts of it.

Importing the entire database in one step

From the Workbench, there is a direct import of the whole database (see figure 7.3):

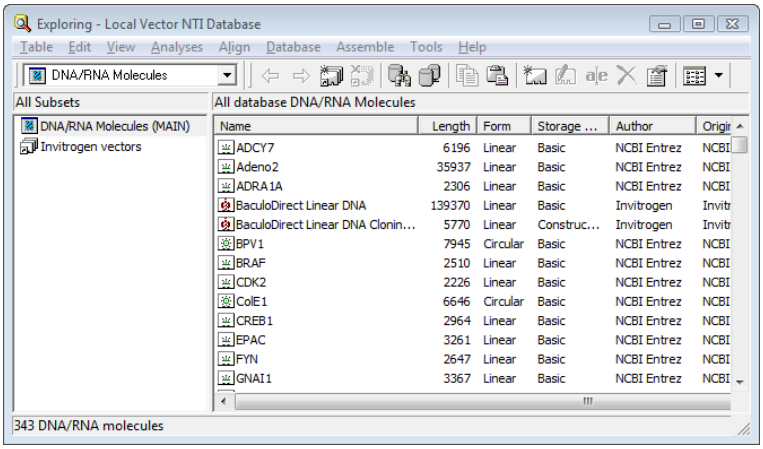

Exploring - Local Vector NTI Database
Table Edit View Analyses Align Database Assemble Tools Help
DNA/RNA Molecules
All Subsets
DNA/RNA Molecules (MAIN)
Invitrogen vectors
All database DNA/RNA Molecules
Name Length Form Storage ... Author Origi
ADCY7 6196 Linear Basic NCBI Entrez NCBI
Adeno2 35937 Linear Basic NCBI Entrez NCBI
ADRA1A 2306 Linear Basic NCBI Entrez NCBI
BaculoDirect Linear DNA 139370 Linear Basic Invitrogen Invitr
BaculoDirect Linear DNA Clonin... 5770 Linear Construc... Invitrogen Invitr
BPV1 7945 Circular Basic NCBI Entrez NCBI
BRAF 2510 Linear Basic NCBI Entrez NCBI
CDK2 2226 Linear Basic NCBI Entrez NCBI
ColE1 6646 Circular Basic NCBI Entrez NCBI
CREB1 2964 Linear Basic NCBI Entrez NCBI
EPAC 3261 Linear Basic NCBI Entrez NCBI
FYN 2647 Linear Basic NCBI Entrez NCBI
GNAI1 3367 Linear Basic NCBI Entrez NCBI
343 DNA/RNA molecules

Figure 7.2: *Data stored in the Vector NTI Local Database accessed through Vector NTI Explorer.*

File | Import Vector NTI Database

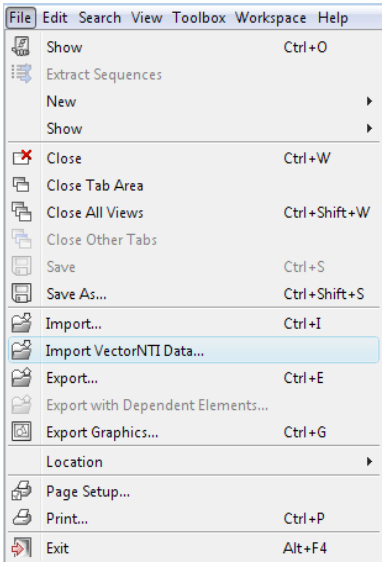


Figure 7.3: *Import the whole Vector NTI Database.*

This will bring up a dialog letting you choose to import from the default location of the database, or you can specify another location. If the database is installed in the default folder, like e.g. *C:\VNTI Database*, press **Yes**. If not, click **No** and specify the database folder manually.

When the import has finished, the data will be listed in the **Navigation Area** of the Workbench as shown in figure 7.4.

If something goes wrong during the import process, please report the problem to support@clcbio.com. To circumvent the problem, see the following section on how to import parts of the database. It will take a few more steps, but you will most likely be able to import this way.

Importing parts of the database

Instead of importing the whole database automatically, you can export parts of the database from Vector NTI Explorer and subsequently import into the Workbench. First, export a selection

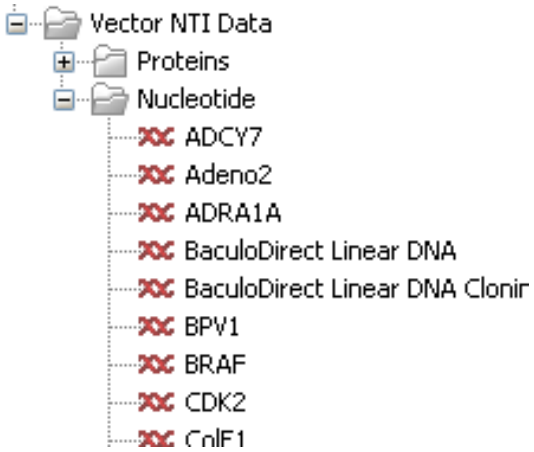


Figure 7.4: *The Vector NTI Data folder containing all imported sequences of the Vector NTI Database.*

of files as an archive as shown in figure 7.5.

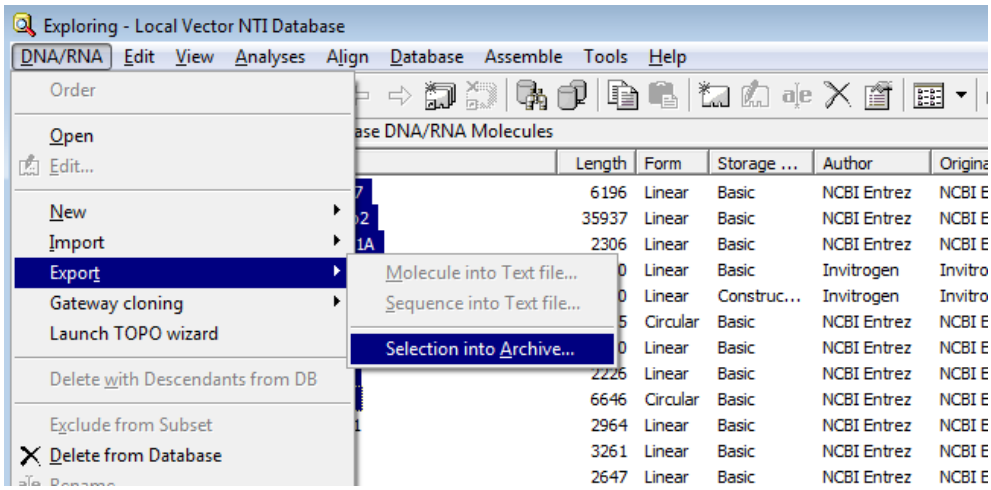


Figure 7.5: *Select the relevant files and export them as an archive through the File menu.*

This will produce a file with a ma4-, pa4- or oa4-extension. Back in the CLC Workbench, click **Import** ( ) and select the file.

Importing single files

In Vector NTI, you can save a sequence in a file instead of in the database (see figure 7.6).

This will give you file with a .gb extension. This file can be easily imported into the CLC Workbench:

Import ( ) | select the file | Select

You don't have to import one file at a time. You can simply select a bunch of files or an entire folder, and the CLC Workbench will take care of the rest. Even if the files are in different formats.

You can also simply drag and drop the files into the **Navigation Area** of the CLC Workbench.

The Vector NTI import is a plug-in which is pre-installed in the Workbench. It can be uninstalled and updated using the plug-in manager (see section 1.7).

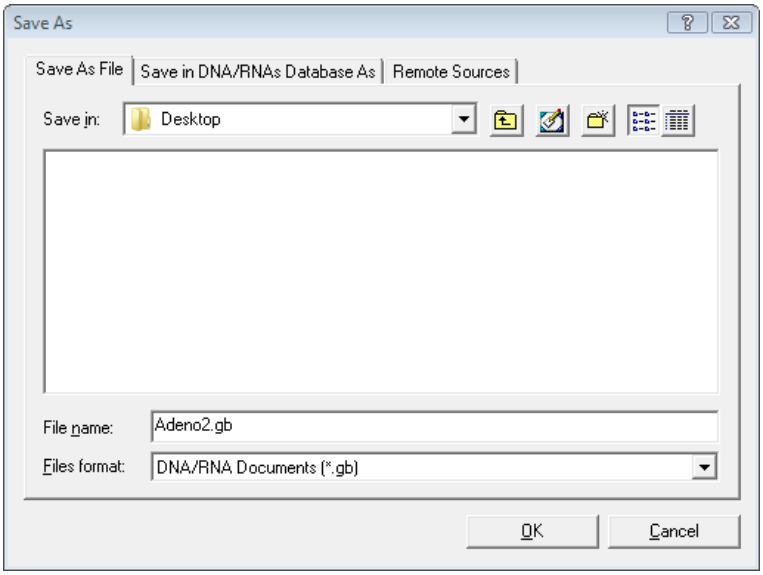


Figure 7.6: *Saving a sequence as a file in Vector NTI.*

7.1.3 Export of bioinformatics data

CLC Protein Workbench can export bioinformatic data in most of the formats that can be imported. There are a few exceptions. See section 7.1.1.

To export a file:

> **select the element to export** | **Export** ( ) | **choose where to export to** | **select 'File of type'** | **enter name of file** | **Save**

When exporting to CSV and tab delimited files, decimal numbers are formatted according to the Locale setting of the Workbench (see section 5.1). If you open the CSV or tab delimited file with spreadsheet software like Excel, you should make sure that both the Workbench and the spreadsheet software are using the same Locale.

Note! The **Export** dialog decides which types of files you are allowed to export into, depending on what type of data you want to export. E.g. protein sequences can be exported into GenBank, Fasta, Swiss-Prot and CLC-formats.

Export of folders and multiple elements

The .zip file type can be used to export all kinds of files and is therefore especially useful in these situations:

- Export of one or more folders including all underlying elements and folders.
- If you want to export two or more elements into one file.

Export of folders is similar to export of single files. Exporting multiple files (of different formats) is done in .zip-format. This is how you export a folder:

> **select the folder to export** | **Export** ( ) | **choose where to export to** | **enter name** | **Save**

You can export multiple files of the same type into formats other than ZIP (.zip). E.g. two DNA sequences can be exported in GenBank format:

> **select the two sequences by <Ctrl>-click (⌘-click on Mac) or <Shift>-click | Export ( ) | choose where to export to | choose GenBank (.gbk) format | enter name the new file | Save**

Export of dependent elements

When exporting e.g. an alignment, *CLC Protein Workbench* can export the alignment including all the sequences that were used to create it. This way, when sending your alignment (with the dependent sequences), your colleagues can reproduce your findings with adjusted parameters, if desired. To export with dependent files:

> **select the element in Navigation Area | File in Menu Bar | Export with Dependent Elements | enter name of of the new file | choose where to export to | Save**

The result is a folder containing the exported file with dependent elements, stored automatically in a folder on the desired location of your desk.

Export history

To export an element's history:

> **select the element in Navigation Area Export () | select History PDF(.pdf) | choose where to export to | Save**

The entire history of the element is then exported in pdf format.

The CLC format

CLC Protein Workbench keeps all bioinformatic data in the CLC format. Compared to other formats, the CLC format contains more information about the object, like its history and comments. The CLC format is also able to hold several elements of different types (e.g. an alignment, a graph and a phylogenetic tree). This means that if you are exporting your data to another CLC Workbench, you can use the CLC format to export several elements in one file, and you will preserve all the information.

Note! CLC files can be exported from and imported into all the different CLC Workbenches.

Backup

If you wish to secure your data from computer breakdowns, it is advisable to perform regular backups of your data. Backing up data in the *CLC Protein Workbench* is done in two ways:

- Making a backup of each of the folders represented by the locations in the **Navigation Area**.
- Selecting all locations in the **Navigation Area** and export () in .zip format. The resulting file will contain all the data stored in the **Navigation Area** and can be imported into *CLC Protein Workbench* if you wish to restore from the back-up at some point.

No matter which method is used for backup, you may have to re-define the locations in the **Navigation Area** if you restore your data from a computer breakdown.

7.2 External files

In order to help you organize your research projects, *CLC Protein Workbench* lets you import all kinds of files. E.g. if you have Word, Excel or pdf-files related to your project, you can import them into the **Navigation Area** of *CLC Protein Workbench*. Importing an external file creates a copy of the file which is stored at the location you have chosen for import. The file can now be opened by double-clicking the file in the **Navigation Area**. The file is opened using the default application for this file type (e.g. Microsoft Word for .doc-files and Adobe Reader for .pdf).

External files are imported and exported in the same way as bioinformatics files (see section 7.1.1). Bioinformatics files not recognized by *CLC Protein Workbench* are also treated as external files.

7.3 Export graphics to files

CLC Protein Workbench supports export of graphics into a number of formats. This way, the visible output of your work can easily be saved and used in presentations, reports etc. The **Export Graphics** function ( ) is found in the **Toolbar**.

CLC Protein Workbench uses a WYSIWYG principle for graphics export: What You See Is What You Get. This means that you should use the options in the Side Panel to change how your data, e.g. a sequence, looks in the program. When you export it, the graphics file will look exactly the same way.

It is not possible to export graphics of elements directly from the **Navigation Area**. They must first be opened in a view in order to be exported. To export graphics of the contents of a view:

select tab of View | Graphics ( ) on Toolbar

This will display the dialog shown in figure 7.7.

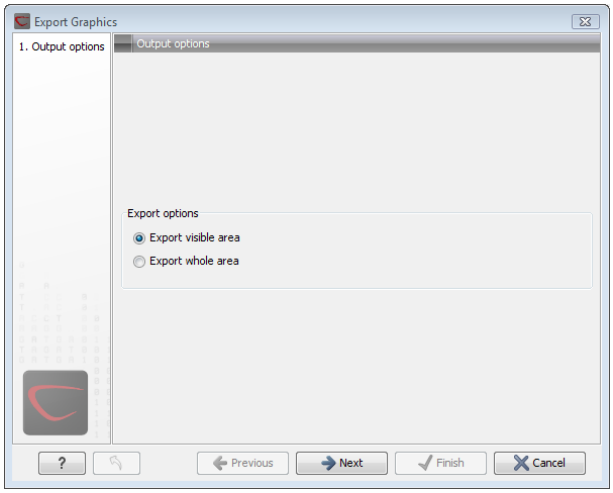


Figure 7.7: *Selecting to export whole view or to export only the visible area.*

7.3.1 Which part of the view to export

In this dialog you can choose to:

- **Export visible area**, or
- **Export whole view**

These options are available for all views that can be zoomed in and out. In figure 7.8 is a view of a circular sequence which is zoomed in so that you can only see a part of it.

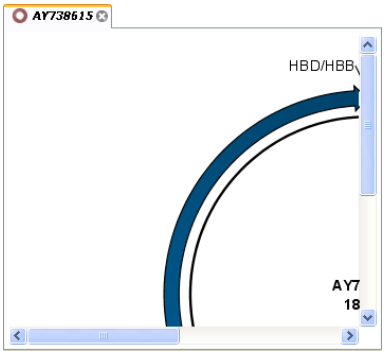


Figure 7.8: *A circular sequence as it looks on the screen.*

When selecting **Export visible area**, the exported file will only contain the part of the sequence that is *visible* in the view. The result from exporting the view from figure 7.8 and choosing **Export visible area** can be seen in figure 7.9.

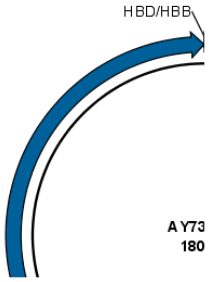


Figure 7.9: *The exported graphics file when selecting Export visible area.*

On the other hand, if you select **Export whole view**, you will get a result that looks like figure 7.10. This means that the graphics file will also include the part of the sequence which is not visible when you have zoomed in.

For 3D structures, this first step is omitted and you will always export what is shown in the view (equivalent to selecting **Export visible area**).

Click **Next** when you have chosen which part of the view to export.

7.3.2 Save location and file formats

In this step, you can choose name and save location for the graphics file (see figure 7.11).

CLC Protein Workbench supports the following file formats for graphics export:

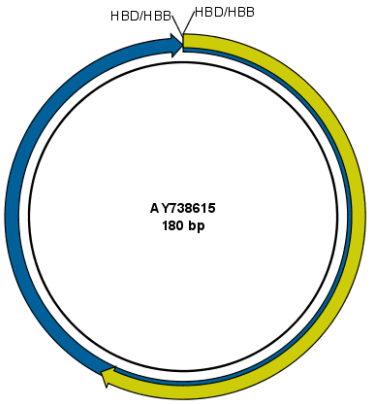


Figure 7.10: *The exported graphics file when selecting Export whole view. The whole sequence is shown, even though the view is zoomed in on a part of the sequence.*

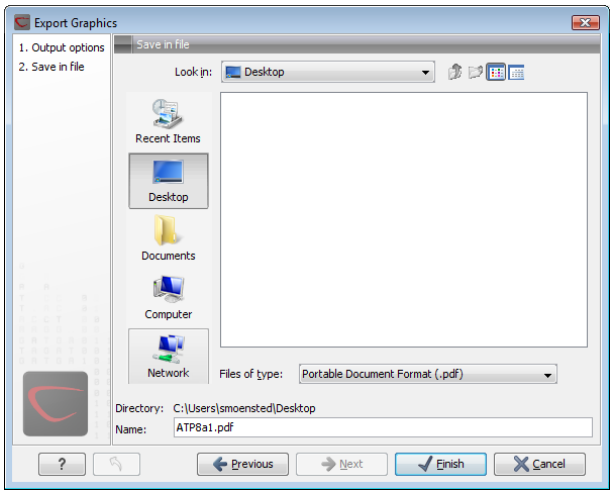


Figure 7.11: *Location and name for the graphics file.*

Format	Suffix	Type
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

These formats can be divided into bitmap and vector graphics. The difference between these two categories is described below:

Bitmap images

In a bitmap image, each dot in the image has a specified color. This implies, that if you zoom in on the image there will not be enough dots, and if you zoom out there will be too many. In these cases the image viewer has to interpolate the colors to fit what is actually looked at. A bitmap image needs to have a high resolution if you want to zoom in. This format is a good choice for storing images without large shapes (e.g. dot plots). It is also appropriate if you don't have the need for resizing and editing the image after export.

Vector graphics

Vector graphic is a collection of shapes. Thus what is stored is e.g. information about where a line starts and ends, and the color of the line and its width. This enables a given viewer to decide how to draw the line, no matter what the zoom factor is, thereby always giving a correct image. This format is good for e.g. graphs and reports, but less usable for e.g. dot plots. If the image is to be resized or edited, vector graphics are by far the best format to store graphics. If you open a vector graphics file in an application like e.g. Adobe Illustrator, you will be able to manipulate the image in great detail.

Graphics files can also be imported into the **Navigation Area**. However, no kinds of graphics files can be displayed in *CLC Protein Workbench*. See section 7.2 for more about importing external files into *CLC Protein Workbench*.

7.3.3 Graphics export parameters

When you have specified the name and location to save the graphics file, you can either click **Next** or **Finish**. Clicking **Next** allows you to set further parameters for the graphics export, whereas clicking **Finish** will export using the parameters that you have set last time you made a graphics export in that file format (if it is the first time, it will use default parameters).

Parameters for bitmap formats

For bitmap files, clicking **Next** will display the dialog shown in figure 7.12.

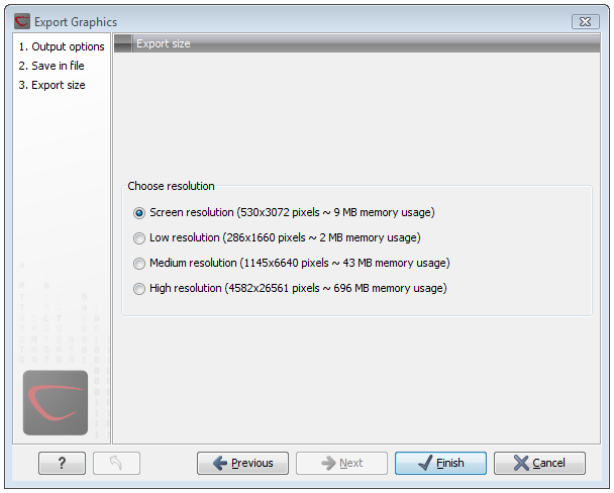

Export Graphics
1. Output options
2. Save in file
3. Export size
Export size
Choose resolution
Screen resolution (530x3072 pixels ~ 9 MB memory usage)
Low resolution (286x1660 pixels ~ 2 MB memory usage)
Medium resolution (1145x6640 pixels ~ 43 MB memory usage)
High resolution (4582x26561 pixels ~ 696 MB memory usage)
Previous
Next
Finish
Cancel

Figure 7.12: *Parameters for bitmap formats: size of the graphics file.*

You can adjust the size (the resolution) of the file to four standard sizes:

- Screen resolution
- Low resolution
- Medium resolution
- High resolution

The actual size in pixels is displayed in parentheses. An estimate of the memory usage for exporting the file is also shown. If the image is to be used on computer screens only, a low resolution is sufficient. If the image is going to be used on printed material, a higher resolution is necessary to produce a good result.

Parameters for vector formats

For pdf format, clicking **Next** will display the dialog shown in figure 7.13 (this is only the case if the graphics is using more than one page).

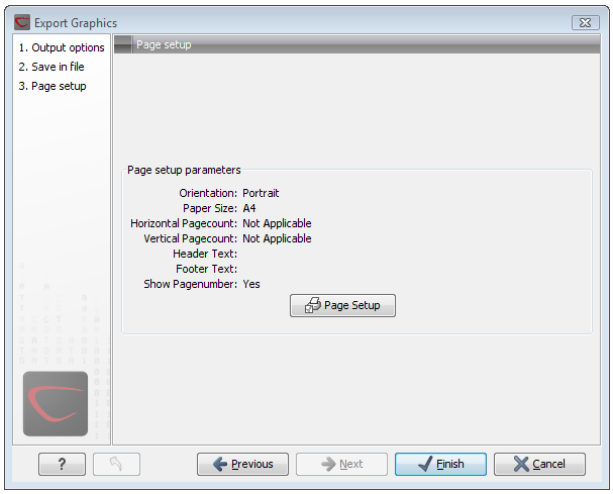


Figure 7.13: *Page setup parameters for vector formats.*

The settings for the page setup are shown, and clicking the **Page Setup** button will display a dialog where these settings can ba adjusted. This dialog is described in section 6.2.

The page setup is only available if you have selected to export the whole view - if you have chosen to export the visible area only, the graphics file will be on one page with no headers or footers.

7.3.4 Exporting protein reports

It is possible to export a protein report using the normal **Export** function ( ) which will generate a pdf file with a table of contents:

Click the report in the Navigation Area | Export ( ) in the Toolbar | select pdf

You can also choose to export a protein report using the **Export graphics** function ( ), but in this way you will not get the table of contents.

7.4 Export graph data points to a file

Data points for graphs displayed along the sequence or along an alignment, mapping or BLAST result, can be exported to a semicolon-separated text file (csv format). An example of such a graph is shown in figure 7.14. This graph shows the coverage of reads of a read mapping (produced with *CLC Genomics Workbench*).

To export the data points for the graph, right-click the graph and choose **Export Graph to Comma-separated File**. Depending on what kind of graph you have selected, different options

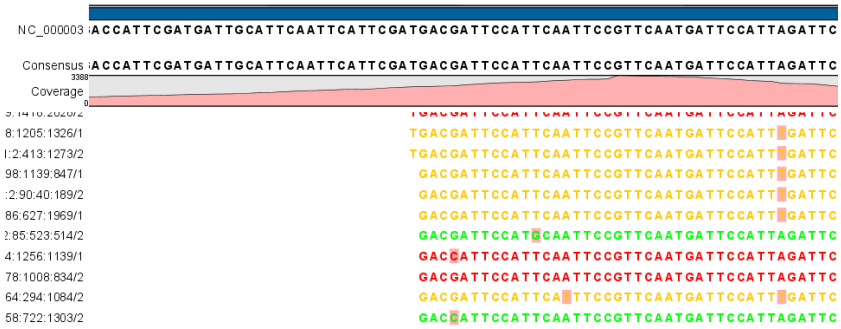


Figure 7.14: *A graph displayed along the mapped reads. Right-click the graph to export the data points to a file.*

will be shown: If the graph is covering a set of aligned sequences with a main sequence, such as read mappings and BLAST results, the dialog shown in figure 7.15 will be displayed. These kinds of graphs are located under **Alignment info** in the Side Panel. In all other cases, a normal file dialog will be shown letting you specify name and location for the file.

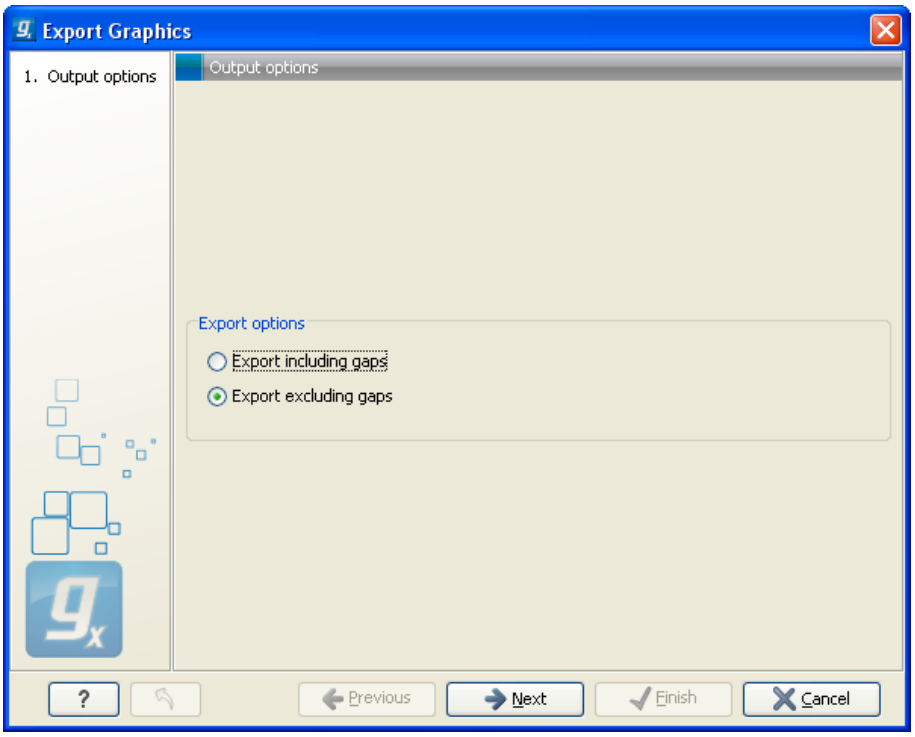


Figure 7.15: *Choosing to include data points with gaps*

In this dialog, select whether you wish to include positions where the main sequence (the reference sequence for read mappings and the query sequence for BLAST results) has gaps. If you are exporting e.g. coverage information from a read mapping, you would probably want to exclude gaps, if you want the positions in the exported file to match the reference (i.e. chromosome) coordinates. If you export including gaps, the data points in the file no longer corresponds to the reference coordinates, because each gap will shift the coordinates.

Clicking **Next** will present a file dialog letting you specify name and location for the file.

The output format of the file is like this:

```
"Position";"Value";
"1";"13";
"2";"16";
"3";"23";
"4";"17";
...
```

7.5 Copy/paste view output

The content of tables, e.g. in reports, folder lists, and sequence lists can be copy/pasted into different programs, where it can be edited. *CLC Protein Workbench* pastes the data in tabulator separated format which is useful if you use programs like Microsoft Word and Excel. There is a huge number of programs in which the copy/paste can be applied. For simplicity, we include one example of the copy/paste function from a **Folder Content** view to Microsoft Excel.

First step is to select the desired elements in the view:

> **click a line in the Folder Content view** | **hold Shift-button** | **press arrow down/up key**

See figure 7.16.

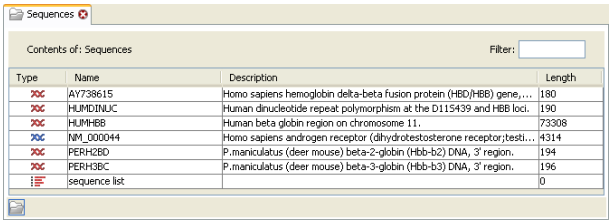


Figure 7.16: *Selected elements in a Folder Content view.*

When the elements are selected, do the following to copy the selected elements:

> **right-click one of the selected elements** | **Edit** | **Copy (** **)**

Then:

> **right-click in the cell A1** | **Paste (** **)**

The outcome might appear unorganized, but with a few operations the structure of the view in *CLC Protein Workbench* can be produced. (Except the icons which are replaced by file references in Excel.)

Note that all tables can also be **Exported** ( ) directly in Excel format.

Chapter 8

History log

Contents

8.1 Element history 116
8.1.1 Sharing data with history 117

CLC Protein Workbench keeps a log of all operations you make in the program. If e.g. you rename a sequence, align sequences, create a phylogenetic tree or translate a sequence, you can always go back and check what you have done. In this way, you are able to document and reproduce previous operations.

This can be useful in several situations: It can be used for documentation purposes, where you can specify exactly how your data has been created and modified. It can also be useful if you return to a project after some time and want to refresh your memory on how the data was created. Also, if you have performed an analysis and you want to reproduce the analysis on another element, you can check the history of the analysis which will give you all parameters you set.

This chapter will describe how to use the **History** functionality of *CLC Protein Workbench*.

8.1 Element history

You can view the history of all elements in the **Navigation Area** except files that are opened in other programs (e.g. Word and pdf-files). The history starts when the element appears for the first time in *CLC Protein Workbench*. To view the history of an element:

Select the element in the Navigation Area | **Show ( ) in the Toolbar** | **History ( )**

or **If the element is already open** | **History ( ) at the bottom left part of the view**

This opens a view that looks like the one in figure 8.1.

When opening an element's history is opened, the newest change is submitted in the top of the view. The following information is available:

- **Title**. The action that the user performed.
- **Date and time**. Date and time for the operation. The date and time are displayed according

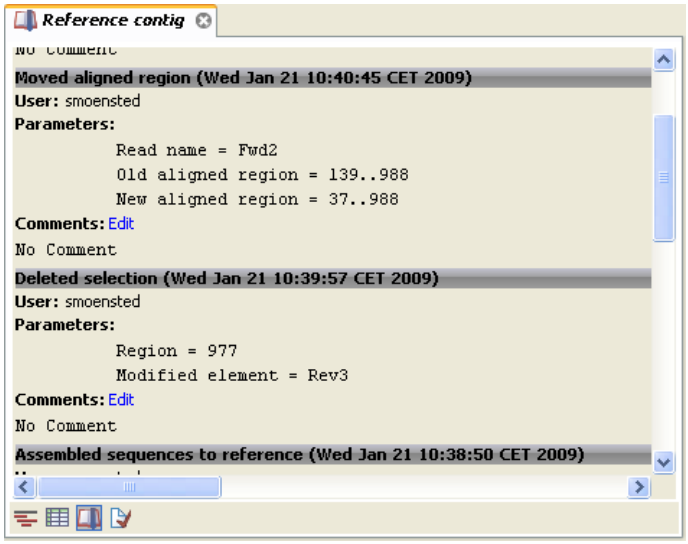


Figure 8.1: *An element's history.*

to your locale settings (see section 5.1).

- **User**. The user who performed the operation. If you import some data created by another person in a CLC Workbench, that persons name will be shown.
- **Parameters**. Details about the action performed. This could be the parameters that was chosen for an analysis.
- **Origins from**. This information is usually shown at the bottom of an element's history. Here, you can see which elements the current element origins from. If you have e.g. created an alignment of three sequences, the three sequences are shown here. Clicking the element selects it in the **Navigation Area**, and clicking the 'history' link opens the element's own history.
- **Comments**. By clicking **Edit** you can enter your own comments regarding this entry in the history. These comments are saved.

8.1.1 Sharing data with history

The history of an element is attached to that element, which means that exporting an element in CLC format (*.clc) will export the history too. In this way, you can share folders and files with others while preserving the history. If an element's history includes source elements (i.e. if there are elements listed in 'Origins from'), they must also be exported in order to see the full history. Otherwise, the history will have entries named "Element deleted". An easy way to export an element with all its source elements is to use the **Export Dependent Elements** function described in section 7.1.3.

The history view can be printed. To do so, click the **Print** icon ( ). The history can also be exported as a pdf file:

Select the element in the Navigation Area | Export ( ) | in "File of type" choose History PDF | Save

Chapter 9

Batching and result handling

Contents

9.1 How to handle results of analyses 118
9.1.1 Table outputs 119
9.1.2 Batch log 120

9.1 How to handle results of analyses

This section will explain how results generated from tools in the Toolbox are handled by *CLC Protein Workbench*. Note that this also applies to tools not running in batch mode (see above). All the analyses in the **Toolbox** are performed in a step-by-step procedure. First, you select elements for analyses, and then there are a number of steps where you can specify parameters (some of the analyses have no parameters, e.g. when translating DNA to RNA). The final step concerns the handling of the results of the analysis, and it is almost identical for all the analyses so we explain it in this section in general.

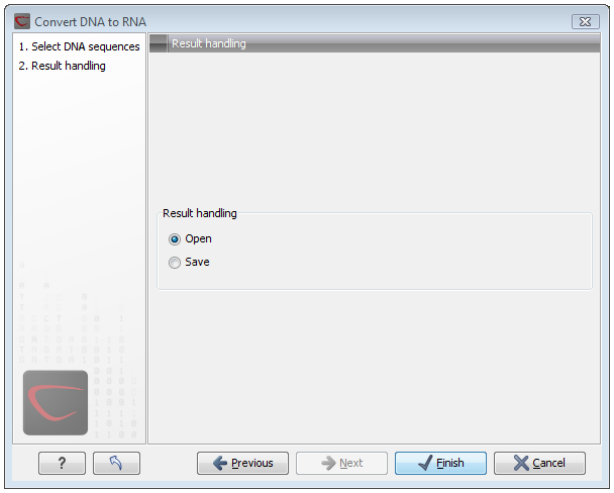


Figure 9.1: *The last step of the analyses exemplified by Translate DNA to RNA.*

In this step, shown in figure 9.1, you have two options:

- **Open.** This will open the result of the analysis in a view. This is the default setting.
- **Save.** This means that the result will not be opened but saved to a folder in the **Navigation Area**. If you select this option, click **Next** and you will see one more step where you can specify where to save the results (see figure 9.2). In this step, you also have the option of creating a new folder or adding a location by clicking the buttons ( )/ ( ) at the top of the dialog.

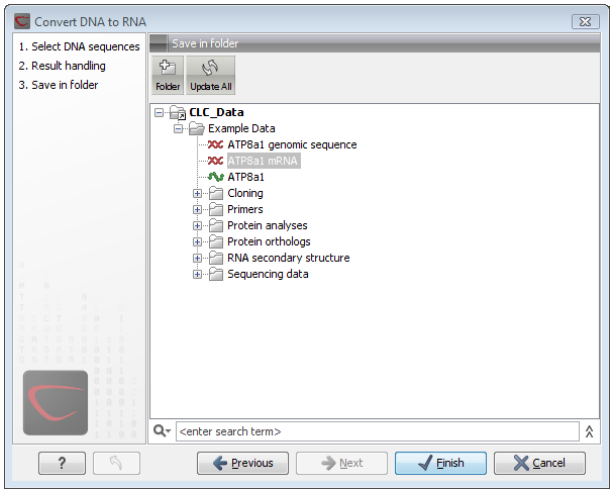


Figure 9.2: *Specify a folder for the results of the analysis.*

9.1.1 Table outputs

Some analyses also generate a table with results, and for these analyses the last step looks like figure 9.3.

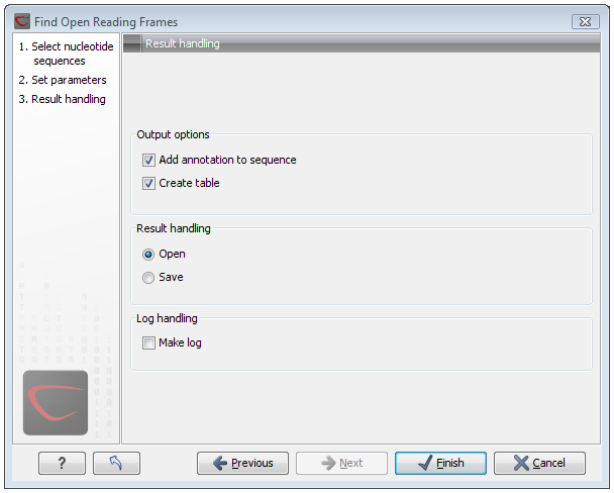


Figure 9.3: *Analyses which also generate tables.*

In addition to the **Open** and **Save** options you can also choose whether the result of the analysis should be added as annotations on the sequence or shown on a table. If both options are selected, you will be able to click the results in the table and the corresponding region on the sequence will be selected.

If you choose to add annotations to the sequence, they can be removed afterwards by clicking **Undo** ( ) in the **Toolbar**.

9.1.2 Batch log

For some analyses, there is an extra option in the final step to create a log of the batch process (see e.g. figure 9.3). This log will be created in the beginning of the process and continually updated with information about the results. See an example of a log in figure 9.4. In this example, the log displays information about how many open reading frames were found.

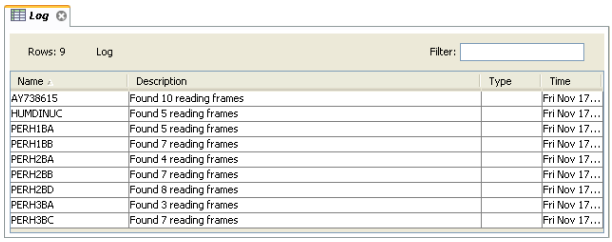

Log
Rows: 9
Log
Filter:
Name
Description
Type
Time
AY738615 Found 10 reading frames Fri Nov 17...
HUMDINUC Found 5 reading frames Fri Nov 17...
PERH1BA Found 5 reading frames Fri Nov 17...
PERH1BB Found 7 reading frames Fri Nov 17...
PERH2BA Found 4 reading frames Fri Nov 17...
PERH2BB Found 7 reading frames Fri Nov 17...
PERH2BD Found 8 reading frames Fri Nov 17...
PERH3BA Found 3 reading frames Fri Nov 17...
PERH3BC Found 7 reading frames Fri Nov 17...

Figure 9.4: *An example of a batch log when finding open reading frames.*

The log will either be saved with the results of the analysis or opened in a view with the results, depending on how you chose to handle the results.

Part III

Bioinformatics

Chapter 10

Viewing and editing sequences

Contents

10.1 View sequence . . . 122
10.1.1 Sequence settings in Side Panel . . . 123
10.1.2 Restriction sites in the Side Panel . . . 129
10.1.3 Selecting parts of the sequence . . . 129
10.1.4 Editing the sequence . . . 130
10.1.5 Sequence region types . . . 130
10.2 Circular DNA . . . 130
10.2.1 Using split views to see details of the circular molecule . . . 132
10.2.2 Mark molecule as circular and specify starting point . . . 132
10.3 Working with annotations . . . 133
10.3.1 Viewing annotations . . . 133
10.3.2 Adding annotations . . . 137
10.3.3 Edit annotations . . . 139
10.3.4 Removing annotations . . . 140
10.4 Element information . . . 141
10.5 View as text . . . 142
10.6 Creating a new sequence . . . 142
10.7 Sequence Lists . . . 143
10.7.1 Graphical view of sequence lists . . . 144
10.7.2 Sequence list table . . . 144
10.7.3 Extract sequences . . . 145

CLC Protein Workbench offers five different ways of viewing and editing single sequences as described in the first five sections of this chapter. Furthermore, this chapter also explains how to create a new sequence and how to gather several sequences in a sequence list.

10.1 View sequence

When you double-click a sequence in the **Navigation Area**, the sequence will open automatically, and you will see the nucleotides or amino acids. The zoom options described in section 3.3 allow

you to e.g. zoom out in order to see more of the sequence in one view. There are a number of options for viewing and editing the sequence which are all described in this section. All the options described in this section also apply to alignments (further described in section 18.2).

10.1.1 Sequence settings in Side Panel

Each view of a sequence has a **Side Panel** located at the right side of the view (see figure 10.1.

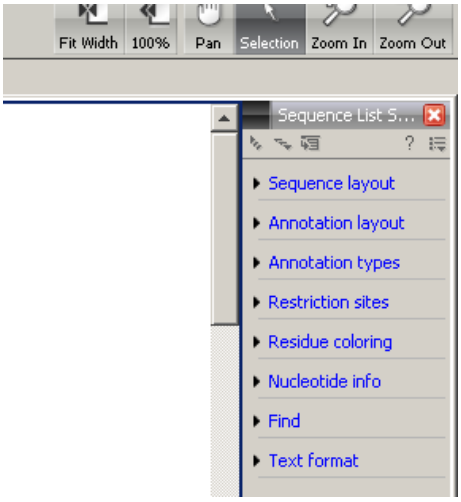


Figure 10.1: *Overview of the Side Panel which is always shown to the right of a view.*

When you make changes in the **Side Panel** the view of the sequence is instantly updated. To show or hide the **Side Panel**:

select the View | Ctrl + U

or **Click the ( ) at the top right corner of the Side Panel to hide | Click the gray Side Panel button to the right to show**

Below, each group of settings will be explained. Some of the preferences are not the same for nucleotide and protein sequences, but the differences will be explained for each group of settings.

Note! When you make changes to the settings in the **Side Panel**, they are not automatically saved when you save the sequence. Click **Save/restore Settings** ( ) to save the settings (see section 5.5 for more information).

Sequence Layout

These preferences determine the overall layout of the sequence:

- **Spacing.** Inserts a space at a specified interval:
  - **No spacing.** The sequence is shown with no spaces.
  - **Every 10 residues.** There is a space every 10 residues, starting from the beginning of the sequence.
  - **Every 3 residues, frame 1.** There is a space every 3 residues, corresponding to the reading frame starting at the first residue.

  - **Every 3 residues, frame 2.** There is a space every 3 residues, corresponding to the reading frame starting at the second residue.
  - **Every 3 residues, frame 3.** There is a space every 3 residues, corresponding to the reading frame starting at the third residue.

- **Wrap sequences.** Shows the sequence on more than one line.
  - **No wrap.** The sequence is displayed on one line.
  - **Auto wrap.** Wraps the sequence to fit the width of the view, not matter if it is zoomed in our out (displays minimum 10 nucleotides on each line).
  - **Fixed wrap.** Makes it possible to specify when the sequence should be wrapped. In the text field below, you can choose the number of residues to display on each line.
- **Double stranded.** Shows both strands of a sequence (only applies to DNA sequences).
- **Numbers on sequences.** Shows residue positions along the sequence. The starting point can be changed by setting the number in the field below. If you set it to e.g. 101, the first residue will have the position of -100. This can also be done by right-clicking an annotation and choosing **Set Numbers Relative to This Annotation**.
- **Numbers on plus strand.** Whether to set the numbers relative to the positive or the negative strand in a nucleotide sequence (only applies to DNA sequences).
- **Follow selection.** When viewing the same sequence in two separate views, "Follow selection" will automatically scroll the view in order to follow a selection made in the other view.
- **Lock numbers.** When you scroll vertically, the position numbers remain visible. (Only possible when the sequence is not wrapped.)
- **Lock labels.** When you scroll horizontally, the label of the sequence remains visible.
- **Sequence label.** Defines the label to the left of the sequence.
  - Name (this is the default information to be shown).
  - Accession (sequences downloaded from databases like GenBank have an accession number).
  - Latin name.
  - Latin name (accession).
  - Common name.
  - Common name (accession).

Annotation Layout and Annotation Types

See section 10.3.1.

Restriction sites

See section 10.1.2.

Motifs

See section 14.7.1.

Residue coloring

These preferences make it possible to color both the residue letter and set a background color for the residue.

- **Non-standard residues.** For nucleotide sequences this will color the residues that are not C, G, A, T or U. For amino acids only B, Z, and X are colored as non-standard residues.
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.
- **Rasmol colors.** Colors the residues according to the Rasmol color scheme.
  See http://www.openrasmol.org/doc/rasmol.html
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.
- **Polarity colors (only protein).** Colors the residues according to the polarity of amino acids.
  - **Foreground color.** Sets the color of the letter. Click the color box to change the color.
  - **Background color.** Sets the background color of the residues. Click the color box to change the color.
- **Trace colors (only DNA).** Colors the residues according to the color conventions of chromatogram traces: A=green, C=blue, G=black, and T=red.
  - **Foreground color.** Sets the color of the letter.
  - **Background color.** Sets the background color of the residues.

Nucleotide info

These preferences only apply to nucleotide sequences.

- **Translation.** Displays a translation into protein just below the nucleotide sequence. Depending on the zoom level, the amino acids are displayed with three letters or one letter.
  - **Frame.** Determines where to start the translation.
    - **ORF/CDS**. If the sequence is annotated, the translation will follow the CDS or ORF annotations. If annotations overlap, only one translation will be shown. If only one annotation is visible, the Workbench will attempt to use this annotation to mark the start and stop for the translation. In cases where this is not possible, the first annotation will be used (i.e. the one closest to the 5' end of the sequence).

        - **Selection.** This option will only take effect when you make a selection on the sequence. The translation will start from the first nucleotide selected. Making a new selection will automatically display the corresponding translation. Read more about selecting in section 10.1.3.
        - **+1 to -1.** Select one of the six reading frames.
        - **All forward/All reverse.** Shows either all forward or all reverse reading frames.
        - **All.** Select all reading frames at once. The translations will be displayed on top of each other.
    - **Table.** The translation table to use in the translation. For more about translation tables, see section 15.5.
    - **Only AUG start codons.** For most genetic codes, a number of codons can be start codons. Selecting this option only colors the AUG codons green.
    - **Single letter codes.** Choose to represent the amino acids with a single letter instead of three letters.

- **G/C content.** Calculates the G/C content of a part of the sequence and shows it as a gradient of colors or as a graph below the sequence.

    - **Window length.** Determines the length of the part of the sequence to calculate. A window length of 9 will calculate the G/C content for the nucleotide in question plus the 4 nucleotides to the left and the 4 nucleotides to the right. A narrow window will focus on small fluctuations in the G/C content level, whereas a wider window will show fluctuations between larger parts of the sequence.
    - **Foreground color.** Colors the letter using a gradient, where the left side color is used for low levels of G/C content and the right side color is used for high levels of G/C content. The sliders just above the gradient color box can be dragged to highlight relevant levels of G/C content. The colors can be changed by clicking the box. This will show a list of gradients to choose from.
    - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
    - **Graph.** The G/C content level is displayed on a graph (Learn how to export the data behind the graph in section 7.4).
        - **Height.** Specifies the height of the graph.
        - **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
        - **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. For Colors, the color box is replaced by a gradient color box as described under Foreground color.

Protein info

These preferences only apply to proteins. The first nine items are different hydrophobicity scales and are described in section 16.5.2.

- **Kyte-Doolittle.** The Kyte-Doolittle scale is widely used for detecting hydrophobic regions in proteins. Regions with a positive value are hydrophobic. This scale can be used for identifying both surface-exposed regions as well as transmembrane regions, depending on the window size used. Short window sizes of 5-7 generally work well for predicting

putative surface-exposed regions. Large window sizes of 19-21 are well suited for finding transmembrane domains if the values calculated are above 1.6 [Kyte and Doolittle, 1982]. These values should be used as a rule of thumb and deviations from the rule may occur.

- **Cornette.** Cornette *et al.* computed an optimal hydrophobicity scale based on 28 published scales [Cornette et al., 1987]. This optimized scale is also suitable for prediction of alpha-helices in proteins.
- **Engelman.** The Engelman hydrophobicity scale, also known as the GES-scale, is another scale which can be used for prediction of protein hydrophobicity [Engelman et al., 1986]. As the Kyte-Doolittle scale, this scale is useful for predicting transmembrane regions in proteins.
- **Eisenberg.** The Eisenberg scale is a normalized consensus hydrophobicity scale which shares many features with the other hydrophobicity scales [Eisenberg et al., 1984].
- **Rose.** The hydrophobicity scale by Rose *et al.* is correlated to the average area of buried amino acids in globular proteins [Rose et al., 1985]. This results in a scale which is not showing the helices of a protein, but rather the surface accessibility.
- **Janin.** This scale also provides information about the accessible and buried amino acid residues of globular proteins [Janin, 1979].
- **Hopp-Woods.** Hopp and Woods developed their hydrophobicity scale for identification of potentially antigenic sites in proteins. This scale is basically a hydrophilic index where apolar residues have been assigned negative values. Antigenic sites are likely to be predicted when using a window size of 7 [Hopp and Woods, 1983].
- **Welling.** [Welling et al., 1985] Welling et al. used information on the relative occurrence of amino acids in antigenic regions to make a scale which is useful for prediction of antigenic regions. This method is better than the Hopp-Woods scale of hydrophobicity which is also used to identify antigenic regions.
- **Kolaskar-Tongaonkar.** A semi-empirical method for prediction of antigenic regions has been developed [Kolaskar and Tongaonkar, 1990]. This method also includes information of surface accessibility and flexibility and at the time of publication the method was able to predict antigenic determinants with an accuracy of 75%.
- **Surface Probability.** Display of surface probability based on the algorithm by [Emini et al., 1985]. This algorithm has been used to identify antigenic determinants on the surface of proteins.
- **Chain Flexibility.** Display of backbone chain flexibility based on the algorithm by [Karplus and Schulz, 1985]. It is known that chain flexibility is an indication of a putative antigenic determinant.

Find

The Find function can also be invoked by pressing Ctrl + Shift + F (⌘ + Shift + F on Mac).

The Find function can be used for searching the sequence. Clicking the find button will search for the first occurrence of the search term. Clicking the find button again will find the next occurrence and so on. If the search string is found, the corresponding part of the sequence will be selected.

- **Search term.** Enter the text to search for. The search function does not discriminate between lower and upper case characters.
- **Sequence search.** Search the nucleotides or amino acids. For amino acids, the single letter abbreviations should be used for searching. The sequence search also has a set of advanced search parameters:
  - Include negative strand. This will search on the negative strand as well.
  - Treat ambiguous characters as wildcards in search term. If you search for e.g. ATN, you will find both ATG and ATC. If you wish to find literally exact matches for ATN (i.e. only find ATN - not ATG), this option should not be selected.
  - Treat ambiguous characters as wildcards in sequence. If you search for e.g. ATG, you will find both ATG and ATN. If you have large regions of Ns, this option should not be selected.

  Note that if you enter a position instead of a sequence, it will automatically switch to position search.
- **Annotation search.** Searches the annotations on the sequence. The search is performed both on the labels of the annotations, but also on the text appearing in the tooltip that you see when you keep the mouse cursor fixed. If the search term is found, the part of the sequence corresponding to the matching annotation is selected. Below this option you can choose to search for translations as well. Sequences annotated with coding regions often have the translation specified which can lead to undesired results.
- **Position search.** Finds a specific position on the sequence. In order to find an interval, e.g. from position 500 to 570, enter "500..570" in the search field. This will make a selection from position 500 to 570 (both included). Notice the two periods (..) between the start an end number (see section 10.3.2). If you enter positions including thousands separators like 123,345, the comma will just be ignored and it would be equivalent to entering 123345.
- **Include negative strand.** When searching the sequence for nucleotides or amino acids, you can search on both strands.
- **Name search.** Searches for sequence names. This is useful for searching sequence lists, mapping results and BLAST results.

This concludes the description of the **View Preferences**. Next, the options for selecting and editing sequences are described.

Text format

These preferences allow you to adjust the format of all the text in the view (both residue letters, sequence name and translations if they are shown).

- **Text size.** Five different sizes.
- **Font.** Shows a list of Fonts available on your computer.
- **Bold residues.** Makes the residues bold.

10.1.2 Restriction sites in the Side Panel

Please see section 17.1.

10.1.3 Selecting parts of the sequence

You can select parts of a sequence:

Click Selection ( ) in Toolbar | Press and hold down the mouse button on the sequence where you want the selection to start | move the mouse to the end of the selection while holding the button | release the mouse button

Alternatively, you can search for a specific interval using the find function described above.

If you have made a selection and wish to adjust it:

drag the edge of the selection (you can see the mouse cursor change to a horizontal arrow

or **press and hold the Shift key while using the right and left arrow keys to adjust the right side of the selection.**

If you wish to select the entire sequence:

double-click the sequence name to the left

Selecting several parts at the same time (multiselect)

You can select several parts of sequence by holding down the **Ctrl** button while making selections. Holding down the **Shift** button lets you extend or reduce an existing selection to the position you clicked.

To select a part of a sequence covered by an annotation:

right-click the annotation | Select annotation

or **double-click the annotation**

To select a fragment between two restriction sites that are shown on the sequence:

double-click the sequence between the two restriction sites

(Read more about restriction sites in section 10.1.2.)

Open a selection in a new view

A selection can be opened in a new view and saved as a new sequence:

right-click the selection | Open selection in New View ( )

This opens the annotated part of the sequence in a new view. The new sequence can be saved by dragging the tab of the sequence view into the **Navigation Area**.

The process described above is also the way to manually translate coding parts of sequences (CDS) into protein. You simply translate the new sequence into protein. This is done by:

right-click the tab of the new sequence | Toolbox | Nucleotide Analyses ( ) | Translate to Protein ( )

A selection can also be copied to the clipboard and pasted into another program:

make a selection | Ctrl + C (⌘ + C on Mac)

Note! The annotations covering the selection will not be copied.

A selection of a sequence can be edited as described in the following section.

10.1.4 Editing the sequence

When you make a selection, it can be edited by:

right-click the selection | Edit Selection ( )

A dialog appears displaying the sequence. You can add, remove or change the text and click **OK**. The original selected part of the sequence is now replaced by the sequence entered in the dialog. This dialog also allows you to paste text into the sequence using Ctrl + V (⌘ + V on Mac).

If you delete the text in the dialog and press **OK**, the selected text on the sequence will also be deleted. Another way to delete a part of the sequence is to:

right-click the selection | Delete Selection ( )

If you wish to only correct only one residue, this is possible by simply making the selection only cover one residue and then type the new residue.

10.1.5 Sequence region types

The various annotations on sequences cover parts of the sequence. Some cover an interval, some cover intervals with unknown endpoints, some cover more than one interval etc. In the following, all of these will be referred to as *regions*. Regions are generally illustrated by markings (often arrows) on the sequences. An arrow pointing to the right indicates that the corresponding region is located on the positive strand of the sequence. Figure 10.2 is an example of three regions with separate colors.

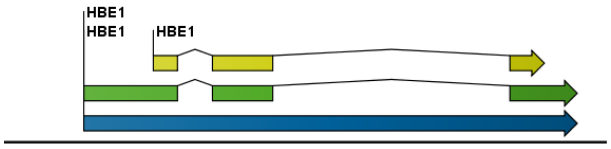


Figure 10.2: *Three regions on a human beta globin DNA sequence (HUMHBB).*

Figure 10.3 shows an artificial sequence with all the different kinds of regions.

10.2 Circular DNA

A sequence can be shown as a circular molecule:

select a sequence in the Navigation Area | Show in the Toolbar | As Circular ( )

or **If the sequence is already open | Click Show As Circular ( ) at the lower left part of the view**

This will open a view of the molecule similar to the one in figure 10.4.

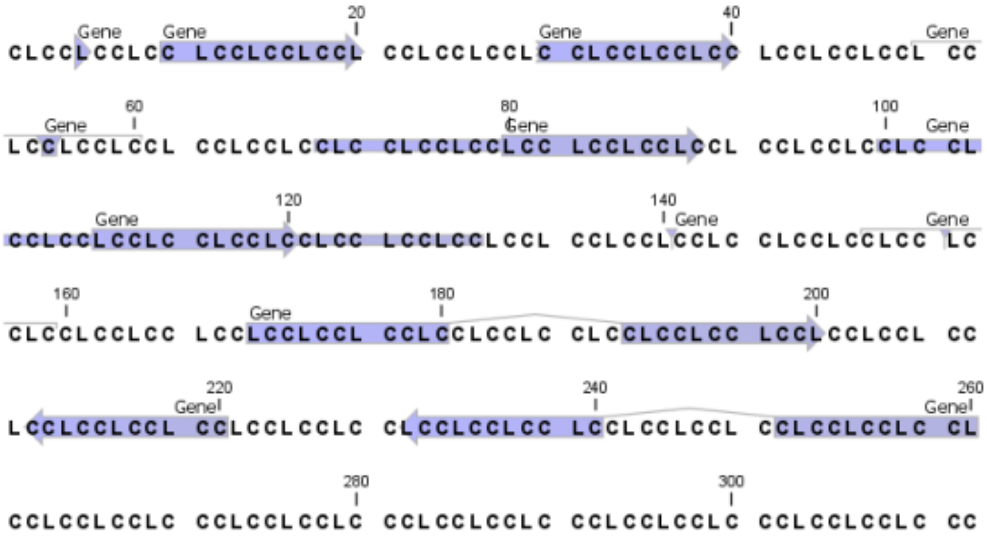


Figure 10.3: *Region #1: A single residue, Region #2: A range of residues including both endpoints, Region #3: A range of residues starting somewhere before 30 and continuing up to and including 40, Region #4: A single residue somewhere between 50 and 60 inclusive, Region #5: A range of residues beginning somewhere between 70 and 80 inclusive and ending at 90 inclusive, Region #6: A range of residues beginning somewhere between 100 and 110 inclusive and ending somewhere between 120 and 130 inclusive, Region #7: A site between residues 140 and 141, Region #8: A site between two residues somewhere between 150 and 160 inclusive, Region #9: A region that covers ranges from 170 to 180 inclusive and 190 to 200 inclusive, Region #10: A region on negative strand that covers ranges from 210 to 220 inclusive, Region #11: A region on negative strand that covers ranges from 230 to 240 inclusive and 250 to 260 inclusive.*

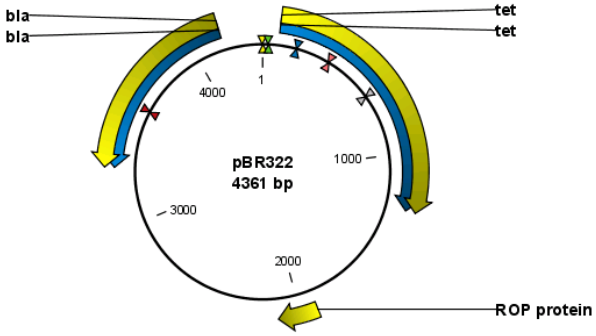


Figure 10.4: *A molecule shown in a circular view.*

This view of the sequence shares some of the properties of the linear view of sequences as described in section 10.1, but there are some differences. The similarities and differences are listed below:

- **Similarities**:
  - The editing options.
  - Options for adding, editing and removing annotations.
  - **Restriction Sites**, **Annotation Types**, **Find** and **Text Format** preferences groups.
- **Differences**:

- In the **Sequence Layout** preferences, only the following options are available in the circular view: **Numbers on plus strand**, **Numbers on sequence** and **Sequence label**.
- You cannot zoom in to see the residues in the circular molecule. If you wish to see these details, split the view with a linear view of the sequence
- In the **Annotation Layout**, you also have the option of showing the labels as **Stacked**. This means that there are no overlapping labels and that all labels of both annotations and restriction sites are adjusted along the left and right edges of the view.

10.2.1 Using split views to see details of the circular molecule

In order to see the nucleotides of a circular molecule you can open a new view displaying a circular view of the molecule:

> **Press and hold the Ctrl button (⌘ on Mac) | click Show Sequence (ACT) at the bottom of the view**

This will open a linear view of the sequence below the circular view. When you zoom in on the linear view you can see the residues as shown in figure 10.5.

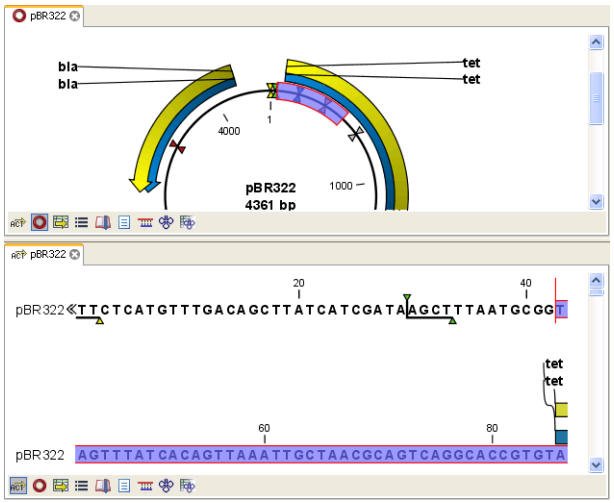


Figure 10.5: *Two views showing the same sequence. The bottom view is zoomed in.*

Note! If you make a selection in one of the views, the other view will also make the corresponding selection, providing an easy way for you to focus on the same region in both views.

10.2.2 Mark molecule as circular and specify starting point

You can mark a DNA molecule as circular by right-clicking its name in either the sequence view or the circular view. In the right-click menu you can also make a circular molecule linear. A circular molecule displayed in the normal sequence view, will have the sequence ends marked with a ».

The starting point of a circular sequence can be changed by:

> **make a selection starting at the position that you want to be the new starting point | right-click the selection | Move Starting Point to Selection Start**

Note! This can only be done for sequence that have been marked as circular.

10.3 Working with annotations

Annotations provide information about specific regions of a sequence. A typical example is the annotation of a gene on a genomic DNA sequence.

Annotations derive from different sources:

- Sequences downloaded from databases like GenBank are annotated.
- In some of the data formats that can be imported into *CLC Protein Workbench*, sequences can have annotations (GenBank, EMBL and Swiss-Prot format).
- The result of a number of analyses in *CLC Protein Workbench* are annotations on the sequence (e.g. finding open reading frames and restriction map analysis).
- You can manually add annotations to a sequence (described in the section 10.3.2).

Note! Annotations are included if you export the sequence in GenBank, Swiss-Prot, EMBL or CLC format. When exporting in other formats, annotations are not preserved in the exported file.

10.3.1 Viewing annotations

Annotations can be viewed in a number of different ways:

- As arrows or boxes in the sequence views:
  - Linear and circular view of sequences ( ) / ( ).
  - Alignments ( ).
  - Graphical view of sequence lists ( ).
  - BLAST views (only the query sequence at the top can have annotations) ( ).
- In the table of annotations ( ).
- In the text view of sequences ()

In the following sections, these view options will be described in more detail.

In all the views except the text view (), annotations can be added, modified and deleted. This is described in the following sections.

View Annotations in sequence views

Figure 10.6 shows an annotation displayed on a sequence.

The various sequence views listed in section 10.3.1 have different default settings for showing annotations. However, they all have two groups in the **Side Panel** in common:

- **Annotation Layout**
- **Annotation Types**

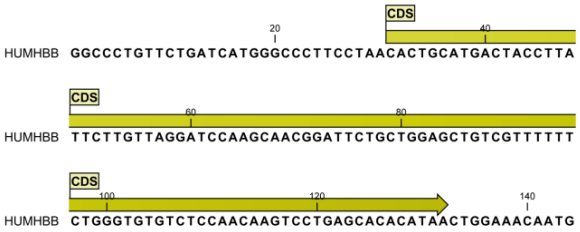


Figure 10.6: *An annotation showing a coding region on a genomic dna sequence.*

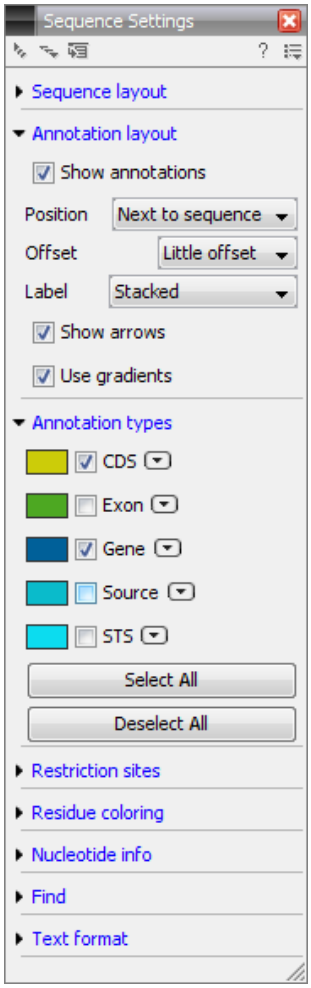


Figure 10.7: *Changing the layout of annotations in the Side Panel.*

The two groups are shown in figure 10.7.

In the **Annotation layout** group, you can specify how the annotations should be displayed (notice that there are some minor differences between the different sequence views):

- **Show annotations.** Determines whether the annotations are shown.
- **Position.**
  - **On sequence.** The annotations are placed on the sequence. The residues are visible through the annotations (if you have zoomed in to 100%).

  - **Next to sequence.** The annotations are placed above the sequence.
  - **Separate layer.** The annotations are placed above the sequence and above restriction sites (only applicable for nucleotide sequences).

- **Offset.** If several annotations cover the same part of a sequence, they can be spread out.
  - **Piled.** The annotations are piled on top of each other. Only the one at front is visible.
  - **Little offset.** The annotations are piled on top of each other, but they have been offset a little.
  - **More offset.** Same as above, but with more spreading.
  - **Most offset.** The annotations are placed above each other with a little space between. This can take up a lot of space on the screen.

- **Label.** The name of the annotation can shown as a label. Additional information about the sequence is shown if you place the mouse cursor on the annotation and keep it still.
  - **No labels.** No labels are displayed.
  - **On annotation.** The labels are displayed in the annotation's box.
  - **Over annotation.** The labels are displayed above the annotations.
  - **Before annotation.** The labels are placed just to the left of the annotation.
  - **Flag.** The labels are displayed as flags at the beginning of the annotation.
  - **Stacked.** The labels are offset so that the text of all labels is visible. This means that there is varying distance between each sequence line to make room for the labels.

- **Show arrows.** Displays the end of the annotation as an arrow. This can be useful to see the orientation of the annotation (for DNA sequences). Annotations on the negative strand will have an arrow pointing to the left.

- **Use gradients.** Fills the boxes with gradient color.

In the **Annotation Types** group, you can choose which kinds of annotations that should be displayed. This group lists all the types of annotations that are attached to the sequence(s) in the view. For sequences with many annotations, it can be easier to get an overview if you deselect the annotation types that are not relevant.

Unchecking the checkboxes in the **Annotation Layout** will not remove this type of annotations them from the sequence - it will just hide them from the view.

Besides selecting which types of annotations that should be displayed, the **Annotation Types** group is also used to change the color of the annotations on the sequence. Click the colored square next to the relevant annotation type to change the color.

This will display a dialog with three tabs: Swatches, HSB, and RGB. They represent three different ways of specifying colors. Apply your settings and click **OK**. When you click **OK**, the color settings cannot be reset. The **Reset** function only works for changes made before pressing **OK**.

Furthermore, the **Annotation Types** can be used to easily browse the annotations by clicking the small button ( ) next to the type. This will display a list of the annotations of that type (see figure 10.8).

Clicking an annotation in the list will select this region on the sequence. In this way, you can quickly find a specific annotation on a long sequence.

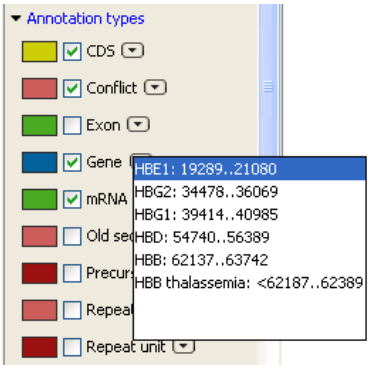


Figure 10.8: *Browsing the gene annotations on a sequence.*

View Annotations in a table

Annotations can also be viewed in a table:

select the sequence in the Navigation Area | Show ( ) | Annotation Table ( )

or **If the sequence is already open | Click Show Annotation Table ( ) at the lower left part of the view**

This will open a view similar to the one in figure 10.9).

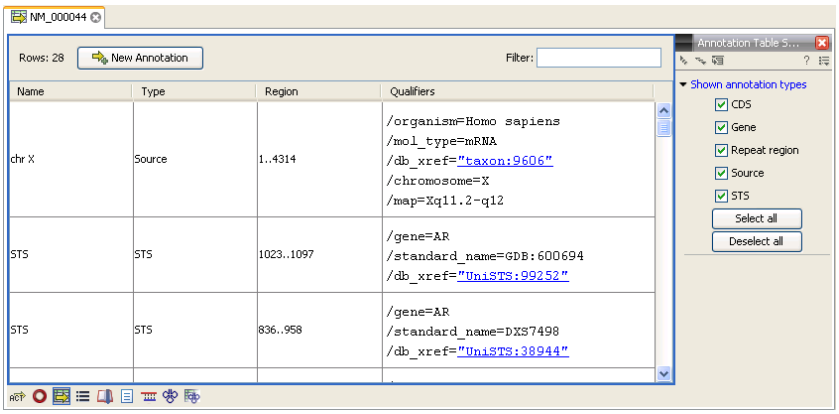


Figure 10.9: *A table showing annotations on the sequence.*

In the **Side Panel** you can show or hide individual annotation types in the table. E.g. if you only wish to see "gene" annotations, de-select the other annotation types so that only "gene" is selected.

Each row in the table is an annotation which is represented with the following information:

- **Name.**
- **Type.**
- **Region.**
- **Qualifiers.**

The Name, Type and Region for each annotation can be edited simply by double-clicking, typing the change directly, and pressing **Enter**.

This information corresponds to the information in the dialog when you edit and add annotations (see section 10.3.2).

You can benefit from this table in several ways:

- It provides an intelligible overview of all the annotations on the sequence.
- You can use the filter at the top to search the annotations. Type e.g. "UCP" into the filter and you will find all annotations which have "UCP" in either the name, the type, the region or the qualifiers. Combined with showing or hiding the annotation types in the **Side Panel**, this makes it easy to find annotations or a subset of annotations.
- You can copy and paste annotations, e.g. from one sequence to another.
- If you wish to edit many annotations consecutively, the double-click editing makes this very fast (see section 10.3.2).

10.3.2 Adding annotations

Adding annotations to a sequence can be done in two ways:

open the sequence in a sequence view (double-click in the Navigation Area) | **make a selection covering the part of the sequence you want to annotate**[1] | **right-click the selection** | **Add Annotation (** **)**

or **select the sequence in the Navigation Area** | **Show (** **)** | **Annotations (** **)** | **Add Annotation (** **)**

This will display a dialog like the one in figure 10.10.

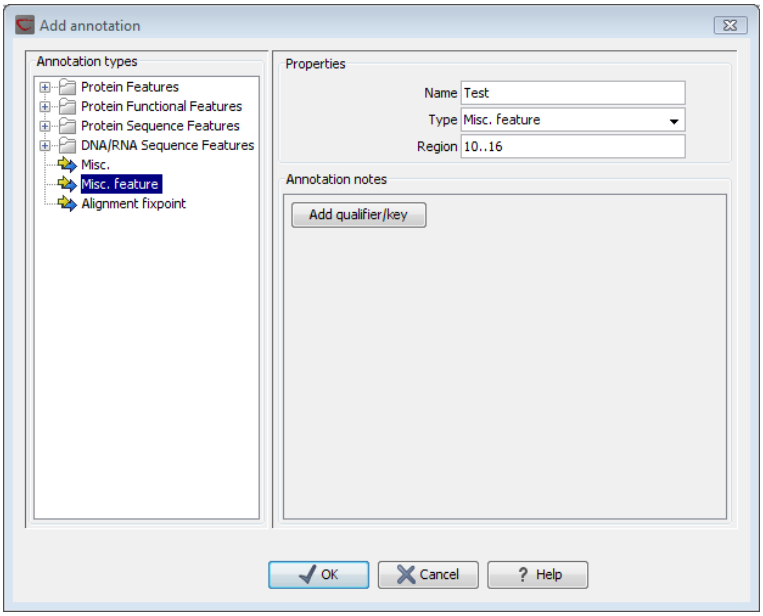


Figure 10.10: *The Add Annotation dialog.*

The left-hand part of the dialog lists a number of **Annotation types**. When you have selected an annotation type, it appears in **Type** to the right. You can also select an annotation directly in this

list. Choosing an annotation type is mandatory. If you wish to use an annotation type which is not present in the list, simply enter this type into the **Type** field [2].

The right-hand part of the dialog contains the following text fields:

- **Name.** The name of the annotation which can be shown on the label in the sequence views. (Whether the name is actually shown depends on the **Annotation Layout** preferences, see section 10.3.1).
- **Type.** Reflects the left-hand part of the dialog as described above. You can also choose directly in this list or type your own annotation type.
- **Region.** If you have already made a selection, this field will show the positions of the selection. You can modify the region further using the conventions of DDBJ, EMBL and GenBank. The following are examples of how to use the syntax (based on http://www.ncbi.nlm.nih.gov/collab/FT/):
  - **467**. Points to a single residue in the presented sequence.
  - **340..565**. Points to a continuous range of residues bounded by and including the starting and ending residues.
  - **<345..500**. Indicates that the exact lower boundary point of a region is unknown. The location begins at some residue previous to the first residue specified (which is not necessarily contained in the presented sequence) and continues up to and including the ending residue.
  - **<1..888**. The region starts before the first sequenced residue and continues up to and including residue 888.
  - **1..>888**. The region starts at the first sequenced residue and continues beyond residue 888.
  - **(102.110)**. Indicates that the exact location is unknown, but that it is one of the residues between residues 102 and 110, inclusive.
  - **123^124**. Points to a site between residues 123 and 124.
  - **join(12..78,134..202)**. Regions 12 to 78 and 134 to 202 should be joined to form one contiguous sequence.
  - **complement(34..126)** Start at the residue complementary to 126 and finish at the residue complementary to residue 34 (the region is on the strand complementary to the presented strand).
  - **complement(join(2691..4571,4918..5163))**. Joins regions 2691 to 4571 and 4918 to 5163, then complements the joined segments (the region is on the strand complementary to the presented strand).
  - **join(complement(4918..5163),complement(2691..4571))**. Complements regions 4918 to 5163 and 2691 to 4571, then joins the complemented segments (the region is on the strand complementary to the presented strand).
- **Annotations.** In this field, you can add more information about the annotation like comments and links. Click the **Add qualifier/key** button to enter information. Select a qualifier which

[2] Note that your own annotation types will be converted to "unsure" when exporting in GenBank format. As long as you use the sequence in CLC format, you own annotation type will be preserved

describes the kind of information you wish to add. If an appropriate qualifier is not present in the list, you can type your own qualifier. The pre-defined qualifiers are derived from the GenBank format. You can add as many qualifier/key lines as you wish by clicking the button. Redundant lines can be removed by clicking the delete icon ( ). The information entered on these lines is shown in the annotation table (see section 10.3.1) and in the yellow box which appears when you place the mouse cursor on the annotation. If you write a hyperlink in the **Key** text field, like e.g. "www.clcbio.com", it will be recognized as a hyperlink. Clicking the link in the annotation table will open a web browser.

Click **OK** to add the annotation.

Note! The annotation will be included if you export the sequence in GenBank, Swiss-Prot or CLC format. When exporting in other formats, annotations are not preserved in the exported file.

10.3.3 Edit annotations

To edit an existing annotation from within a sequence view:

right-click the annotation | Edit Annotation ( )

This will show the same dialog as in figure 10.10, with the exception that some of the fields are filled out depending on how much information the annotation contains.

There is another way of quickly editing annotations which is particularly useful when you wish to edit several annotations.

To edit the information, simply double-click and you will be able to edit e.g. the name or the annotation type. If you wish to edit the qualifiers and double-click in this column, you will see the dialog for editing annotations.

Advanced editing of annotations

Sometimes you end up with annotations which do not have a meaningful name. In that case there is an advanced batch rename functionality:

Open the Annotation Table ( ) | select the annotations that you want to rename | right-click the selection | Advanced Rename

This will bring up the dialog shown in figure 10.11.

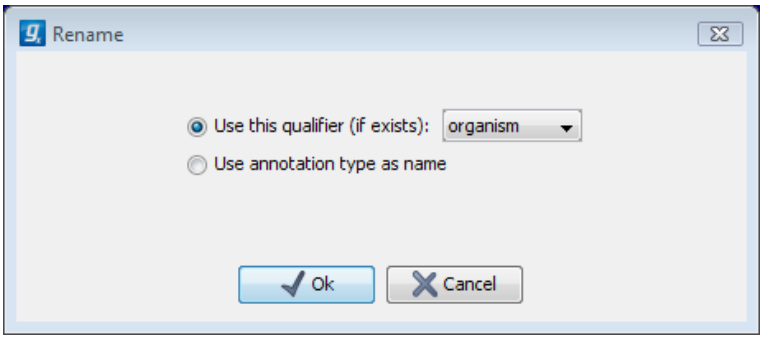


Figure 10.11: *The Advanced Rename dialog.*

In this dialog, you have two options:

- **Use this qualifier.** Use one of the qualifiers as name. A list of all qualifiers of all the selected annotations is shown. Note that if one of the annotations do not have the qualifier you have chosen, it will not be renamed. If an annotation has multiple qualifiers of the same type, the first is used for naming.
- **Use annotation type as name.** The annotation's type will be used as name (e.g. if you have an annotation of type "Promoter", it will get "Promoter" as its name by using this option).

A similar functionality is available for batch re-typing annotations is available in the right-click menu as well, in case your annotations are not typed correctly:

> **Open the Annotation Table ( ) | select the annotations that you want to retype | right-click the selection | Advanced Retype**

This will bring up the dialog shown in figure 10.12.

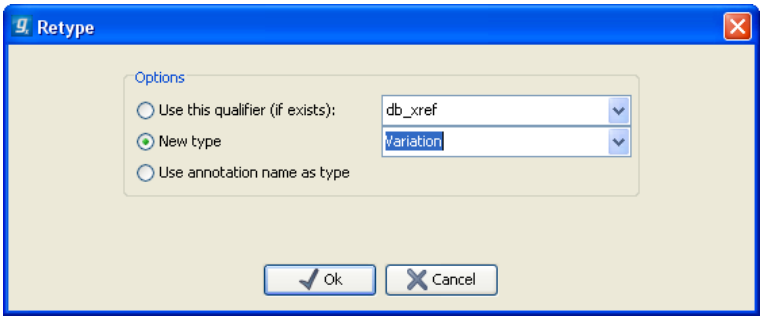


Figure 10.12: *The Advanced Retype dialog.*

In this dialog, you have two options:

- **Use this qualifier.** Use one of the qualifiers as type. A list of all qualifiers of all the selected annotations is shown. Note that if one of the annotations do not have the qualifier you have chosen, it will not be retyped. If an annotation has multiple qualifiers of the same type, the first is used for the new type.
- **New type**. You can select from a list of all the pre-defined types as well as enter your own annotation type. All the selected annotations will then get this type.
- **Use annotation name as type.** The annotation's name will be used as type (e.g. if you have an annotation named "Promoter", it will get "Promoter" as its type by using this option).

10.3.4 Removing annotations

Annotations can be hidden using the **Annotation Types** preferences in the **Side Panel** to the right of the view (see section 10.3.1). In order to completely remove the annotation:

> **right-click the annotation | Delete | Delete Annotation ( )**

If you want to remove all annotations of one type:

> **right-click an annotation of the type you want to remove | Delete | Delete Annotations of Type "type"**

If you want to remove all annotations from a sequence:

right-click an annotation | Delete | Delete All Annotations

The removal of annotations can be undone using Ctrl + Z or Undo ( ) in the Toolbar.

If you have more sequences (e.g. in a sequence list, alignment or contig), you have two additional options:

right-click an annotation | Delete | Delete All Annotations from All Sequences

right-click an annotation | Delete | Delete Annotations of Type "type" from All Sequences

10.4 Element information

The normal view of a sequence (by double-clicking) shows the annotations as boxes along the sequence, but often there is more information available about sequences. This information is available through the **Element info** view.

To view the sequence information:

select a sequence in the Navigation Area | Show ( ) in the Toolbar | Element info ( )

This will display a view similar to fig 10.13.

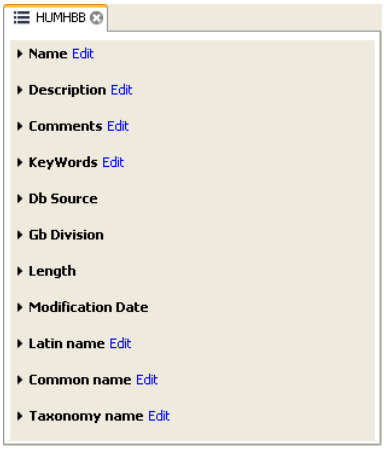


Figure 10.13: *The initial display of sequence info for the HUMHBB DNA sequence from the Example data.*

All the lines in the view are headings, and the corresponding text can be shown by clicking the text.

- **Name.** The name of the sequence which is also shown in sequence views and in the **Navigation Area**.
- **Description.** A description of the sequence.
- **Comments.** The author's comments about the sequence.
- **Keywords.** Keywords describing the sequence.
- **Db source.** Accession numbers in other databases concerning the same sequence.

- **Gb Division.** Abbreviation of GenBank divisions. See section 3.3 in the GenBank release notes for a full list of GenBank divisions.
- **Length.** The length of the sequence.
- **Modification date.** Modification date from the database. This means that this date does not reflect your own changes to the sequence. See the history (section 8) for information about the latest changes to the sequence after it was downloaded from the database.
- **Organism.** Scientific name of the organism (first line) and taxonomic classification levels (second and subsequent lines).

The information available depends on the origin of the sequence. Sequences downloaded from database like NCBI and UniProt (see section 12) have this information. On the other hand, some sequence formats like fasta format do not contain this information.

Some of the information can be edited by clicking the blue **Edit** text. This means that you can add your own information to sequences that do not derive from databases.

Note that for other kinds of data, the **Element info** will only have **Name** and **Description**.

10.5 View as text

A sequence can be viewed as text without any layout and text formatting. This displays all the information about the sequence in the GenBank file format. To view a sequence as text:

select a sequence in the Navigation Area | Show in the Toolbar | As text

This way it is possible to see background information about e.g. the authors and the origin of DNA and protein sequences. Selections or the entire text of the **Sequence Text View** can be copied and pasted into other programs:

Much of the information is also displayed in the **Sequence info**, where it is easier to get an overview (see section 10.4.)

In the **Side Panel**, you find a search field for searching the text in the view.

10.6 Creating a new sequence

A sequence can either be imported, downloaded from an online database or created in the *CLC Protein Workbench*. This section explains how to create a new sequence:

New ( ) in the toolbar

The **Create Sequence** dialog (figure 10.14) reflects the information needed in the GenBank format, but you are free to enter anything into the fields. The following description is a guideline for entering information about a sequence:

- **Name.** The name of the sequence. This is used for saving the sequence.
- **Common name.** A common name for the species.
- **Latin name.** The Latin name for the species.

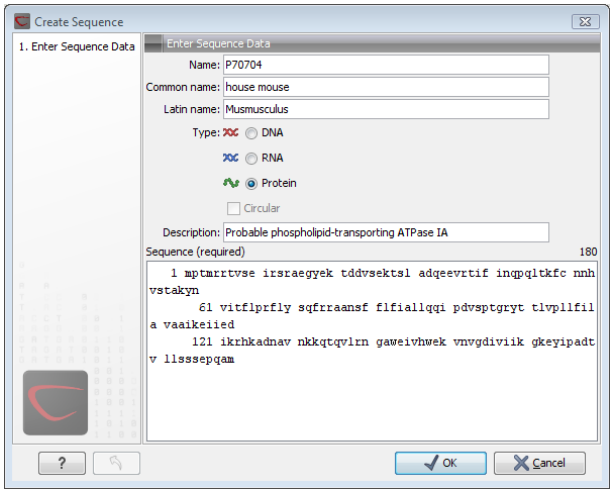


Figure 10.14: *Creating a sequence.*

- **Type.** Select between DNA, RNA and protein.
- **Circular.** Specifies whether the sequence is circular. This will open the sequence in a circular view as default. (applies only to nucleotide sequences).
- **Description.** A description of the sequence.
- **Keywords.** A set of keywords separated by semicolons (;).
- **Comments.** Your own comments to the sequence.
- **Sequence.** Depending on the type chosen, this field accepts nucleotides or amino acids. Spaces and numbers can be entered, but they are ignored when the sequence is created. This allows you to paste (Ctrl + V on Windows and ⌘ + V on Mac) in a sequence directly from a different source, even if the residue numbers are included. Characters that are not part of the IUPAC codes cannot be entered. At the top right corner of the field, the number of residues are counted. The counter does not count spaces or numbers.

Clicking **Finish** opens the sequence. It can be saved by clicking **Save** ( ) or by dragging the tab of the sequence view into the **Navigation Area**.

10.7 Sequence Lists

The **Sequence List** shows a number of sequences in a tabular format or it can show the sequences together in a normal sequence view.

Having sequences in a sequence list can help organizing sequence data. The sequence list may originate from an NCBI search (chapter 11.1). Moreover, if a multiple sequence fasta file is imported, it is possible to store the data in a sequences list. A **Sequence List** can also be generated using a dialog, which is described here:

select two or more sequences | **right-click the elements** | **New** | **Sequence List ( )**

This action opens a **Sequence List** dialog:

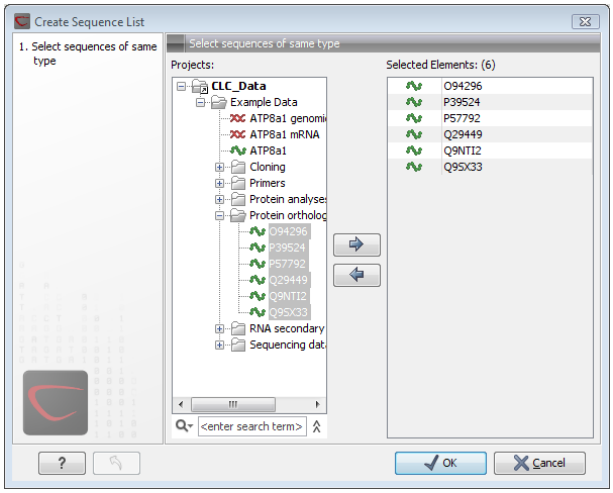


Figure 10.15: *A Sequence List dialog.*

The dialog allows you to select more sequences to include in the list, or to remove already chosen sequences from the list.

Clicking **Finish** opens the sequence list. It can be saved by clicking **Save** ( ) or by dragging the tab of the view into the **Navigation Area**.

Opening a Sequence list is done by:

> **right-click the sequence list in the Navigation Area | Show ( ) | Graphical Sequence List ( ) OR Table ( )**

The two different views of the same sequence list are shown in split screen in figure 10.16.

10.7.1 Graphical view of sequence lists

The graphical view of sequence lists is almost identical to the view of single sequences (see section 10.1). The main difference is that you now can see more than one sequence in the same view.

However, you also have a few extra options for sorting, deleting and adding sequences:

- To add extra sequences to the list, right-click an empty (white) space in the view, and select **Add Sequences**.
- To delete a sequence from the list, right-click the sequence's name and select **Delete Sequence**.
- To sort the sequences in the list, right-click the name of one of the sequences and select **Sort Sequence List by Name** or **Sort Sequence List by Length**.
- To rename a sequence, right-click the name of the sequence and select **Rename Sequence**.

10.7.2 Sequence list table

Each sequence in the table sequence list is displayed with:

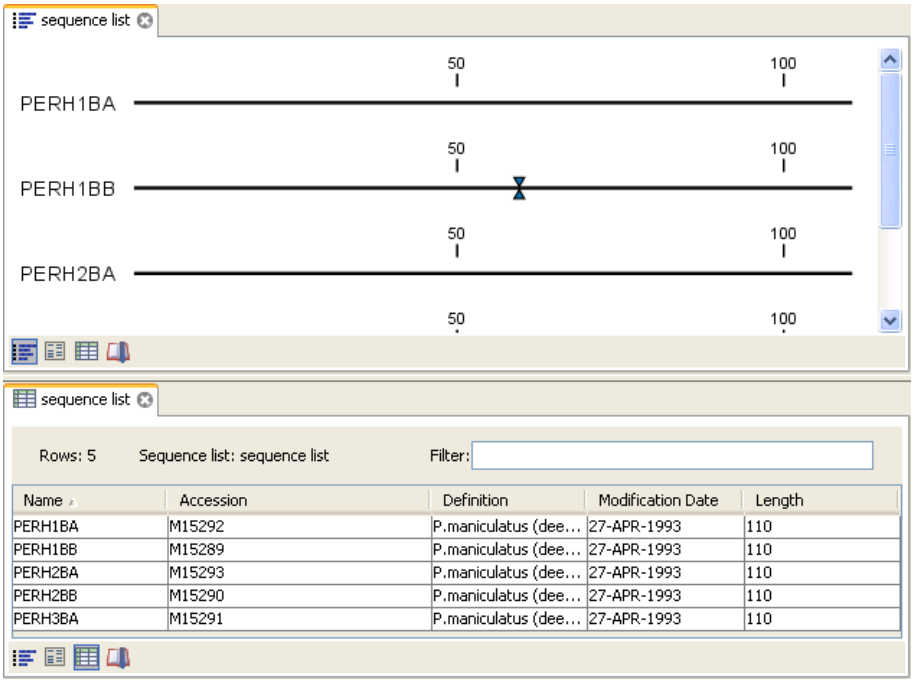


Figure 10.16: *A sequence list containing multiple sequences can be viewed in either a table or in a graphical sequence list. The graphical view is useful for viewing annotations and the sequence itself, while the table view provides other information like sequence lengths, and the number of sequences in the list (number of Rows reported).*

- Name.
- Accession.
- Description.
- Modification date.
- Length.

The number of sequences in the list is reported as the number of Rows at the top of the table view.

Learn more about tables in section C.

Adding and removing sequences from the list is easy: adding is done by dragging the sequence from another list or from the **Navigation Area** and drop it in the table. To delete sequences, simply select them and press **Delete** ( ).

You can also create a subset of the sequence list:

select the relevant sequences | **right-click** | **Create New Sequence List**

This will create a new sequence list which only includes the selected sequences.

10.7.3 Extract sequences

It is possible to extract individual sequences from a sequence list in two ways. If the sequence list is opened in the tabular view, it is possible to drag (with the mouse) one or more sequences

into the **Navigation Area**. This allows you to extract specific sequences from the entire list. Another option is to extract all sequences found in the list. This can also be done for:

- Alignments ( )
- Contigs and read mappings ( )
- Read mapping tables ( )
- BLAST result ( )
- BLAST overview tables ( )
- RNA-Seq samples ( )
- and of course sequence lists ( )

For mappings and BLAST results, the main sequences (i.e. reference/consensus and query sequence) will not be extracted.

To extract the sequences:

Toolbox | **General Sequence Analyses ( )** | **Extract Sequences** ( )

This will allow you to select the elements that you want to extract sequences from (see the list above). Clicking **Next** displays the dialog shown in 10.17.

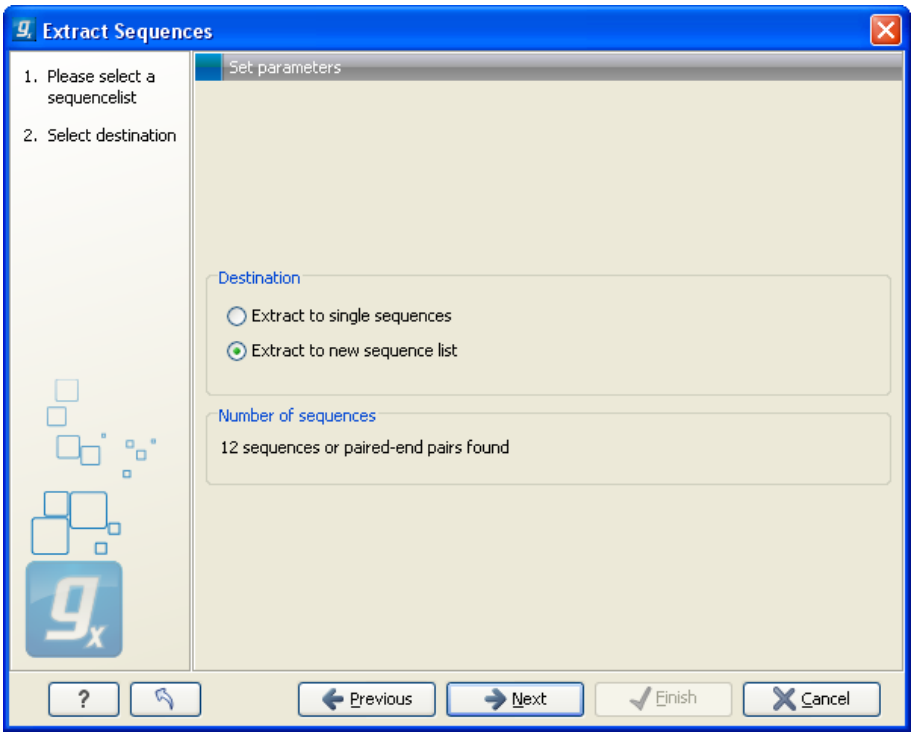


Figure 10.17: *Choosing whether the extracted sequences should be placed in a new list or as single sequences.*

Here you can choose whether the extracted sequences should be placed in a new list or extracted as single sequences. For sequence lists, only the last option makes sense, but for alignments, mappings and BLAST results, it would make sense to place the sequences in a list.

Below these options you can see the number of sequences that will be extracted.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

Chapter 11

Online database search

Contents

11.1 GenBank search 148
11.1.1 GenBank search options 149
11.1.2 Handling of GenBank search results 150
11.1.3 Save GenBank search parameters 151
11.2 UniProt (Swiss-Prot/TrEMBL) search 152
11.2.1 UniProt search options 152
11.2.2 Handling of UniProt search results 153
11.2.3 Save UniProt search parameters 154
11.3 Search for structures at NCBI 154
11.3.1 Structure search options 155
11.3.2 Handling of NCBI structure search results 156
11.3.3 Save structure search parameters 157
11.4 Sequence web info 158
11.4.1 Google sequence 158
11.4.2 NCBI 158
11.4.3 PubMed References 159
11.4.4 UniProt 159
11.4.5 Additional annotation information 159

CLC Protein Workbench offers different ways of searching data on the Internet. You must be online when initiating and performing the following searches:

11.1 GenBank search

This section describes searches for sequences in GenBank - the **NCBI Entrez** database. The NCBI search view is opened in this way (figure 11.1):

Search | Search for Sequences at NCBI ( )

or **Ctrl + B (⌘ + B on Mac)**

This opens the following view:

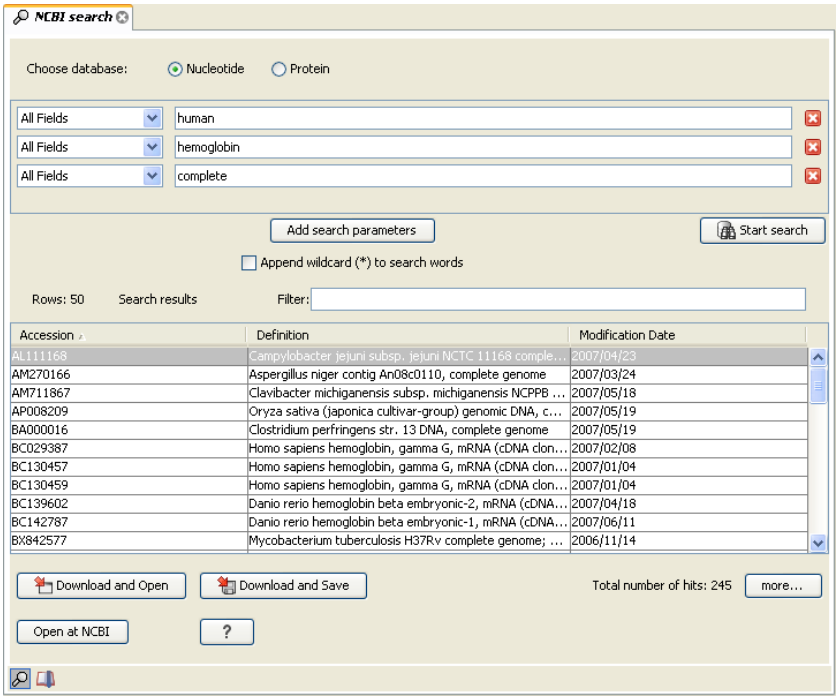

NCBI search
Choose database:
Nucleotide
Protein
All Fields
human
All Fields
hemoglobin
All Fields
complete
Add search parameters
Start search
Append wildcard (*) to search words
Rows: 50
Search results
Filter:
Accession
Definition
Modification Date
AL111168 Campylobacter jejuni subsp. jejuni NCTC 11168 comple... 2007/04/23
AM270166 Aspergillus niger contig An08c0110, complete genome 2007/03/24
AM711867 Clavibacter michiganensis subsp. michiganensis NCPPB ... 2007/05/18
AP008209 Oryza sativa (japonica cultivar-group) genomic DNA, c... 2007/05/19
BA000016 Clostridium perfringens str. 13 DNA, complete genome 2007/05/19
BC029387 Homo sapiens hemoglobin, gamma G, mRNA (cDNA clon... 2007/02/08
BC130457 Homo sapiens hemoglobin, gamma G, mRNA (cDNA clon... 2007/01/04
BC130459 Homo sapiens hemoglobin, gamma G, mRNA (cDNA clon... 2007/01/04
BC139602 Danio rerio hemoglobin beta embryonic-2, mRNA (cDNA... 2007/04/18
BC142787 Danio rerio hemoglobin beta embryonic-1, mRNA (cDNA... 2007/06/11
BX842577 Mycobacterium tuberculosis H37Rv complete genome; ... 2006/11/14
Download and Open
Download and Save
Total number of hits: 245
more...
Open at NCBI
?

Figure 11.1: *The GenBank search view.*

11.1.1 GenBank search options

Conducting a search in the **NCBI Database** from *CLC Protein Workbench* corresponds to conducting the search on NCBI's website. When conducting the search from *CLC Protein Workbench*, the results are available and ready to work with straight away.

You can choose whether you want to search for nucleotide sequences or protein sequences.

As default, *CLC Protein Workbench* offers one text field where the search parameters can be entered. Click **Add search parameters** to add more parameters to your search.
Note! The search is a "and" search, meaning that when adding search parameters to your search, you search for both (or all) text strings rather than "any" of the text strings.

You can append a wildcard character by checking the checkbox at the bottom. This means that you only have to enter the first part of the search text, e.g. searching for "genom" will find both "genomic" and "genome".

The following parameters can be added to the search:

- **All fields**. Text, searches in all parameters in the NCBI database at the same time.
- **Organism**. Text.
- **Description**. Text.
- **Modified Since**. Between 30 days and 10 years.
- **Gene Location**. Genomic DNA/RNA, Mitochondrion, or Chloroplast.
- **Molecule**. Genomic DNA/RNA, mRNA or rRNA.
- **Sequence Length**. Number for maximum or minimum length of the sequence.

- **Gene Name**. Text.

The search parameters are the most recently used. The **All fields** allows searches in all parameters in the NCBI database at the same time. **All fields** also provide an opportunity to restrict a search to parameters which are not listed in the dialog. E.g. writing `gene[Feature key] AND mouse` in **All fields** generates hits in the GenBank database which contains one or more genes and where 'mouse' appears somewhere in GenBank file. You can also write e.g. `CD9 NOT homo sapiens` in **All fields**.

Note! The 'Feature Key' option is only available in GenBank when searching for nucleotide sequences. For more information about how to use this syntax, see http://www.ncbi.nlm.nih.gov/books/NBK3837/

When you are satisfied with the parameters you have entered, click **Start search**.

Note! When conducting a search, no files are downloaded. Instead, the program produces a list of links to the files in the NCBI database. This ensures a much faster search.

11.1.2 Handling of GenBank search results

The search result is presented as a list of links to the files in the NCBI database. The **View** displays 50 hits at a time. This can be changed in the **Preferences** (see chapter 5). More hits can be displayed by clicking the **More...** button at the bottom right of the **View**.

Each sequence hit is represented by text in three columns:

- Accession.
- Description.
- Modification date.
- Length.

It is possible to exclude one or more of these columns by adjust the View preferences for the database search view. Furthermore, your changes in the View preferences can be saved. See section 5.5.

Several sequences can be selected, and by clicking the buttons in the bottom of the search view, you can do the following:

- Download and open, doesn't save the sequence.
- Download and save, lets you choose location for saving sequence.
- Open at NCBI, searches the sequence at NCBI's web page.

Double-clicking a hit will download and open the sequence. The hits can also be copied into the **View Area** or the **Navigation Area** from the search results by drag and drop, copy/paste or by using the right-click menu as described below.

Drag and drop from GenBank search results

The sequences from the search results can be opened by dragging them into a position in the **View Area**.

Note! A sequence is not saved until the **View** displaying the sequence is closed. When that happens, a dialog opens: Save changes of sequence x? (Yes or No).

The sequence can also be saved by dragging it into the **Navigation Area**. It is possible to select more sequences and drag all of them into the **Navigation Area** at the same time.

Download GenBank search results using right-click menu

You may also select one or more sequences from the list and download using the right-click menu (see figure 11.2). Choosing **Download and Save** lets you select a folder where the sequences are saved when they are downloaded. Choosing **Download and Open** opens a new view for each of the selected sequences.

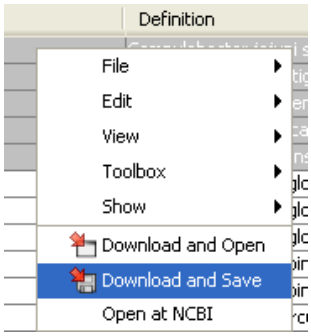


Figure 11.2: *By right-clicking a search result, it is possible to choose how to handle the relevant sequence.*

Copy/paste from GenBank search results

When using copy/paste to bring the search results into the **Navigation Area**, the actual files are downloaded from GenBank.

To copy/paste files into the **Navigation Area**:

> **select one or more of the search results | Ctrl + C (⌘ + C on Mac) | select a folder in the Navigation Area | Ctrl + V**

Note! Search results are downloaded before they are saved. Downloading and saving several files may take some time. However, since the process runs in the background (displayed in the **Status bar**) it is possible to continue other tasks in the program. Like the search process, the download process can be stopped. This is done in the **Toolbox** in the **Processes** tab.

11.1.3 Save GenBank search parameters

The search view can be saved either using dragging the search tab and and dropping it in the **Navigation Area** or by clicking **Save** (). When saving the search, only the parameters are saved - not the results of the search. This is useful if you have a special search that you perform from time to time.

Even if you don't save the search, the next time you open the search view, it will remember the parameters from the last time you did a search.

11.2 UniProt (Swiss-Prot/TrEMBL) search

This section describes searches in UniProt and the handling of search results. UniProt is a global database of protein sequences.

The UniProt search view (figure 11.3) is opened in this way:

Search | Search for Sequences in UniProt ( )

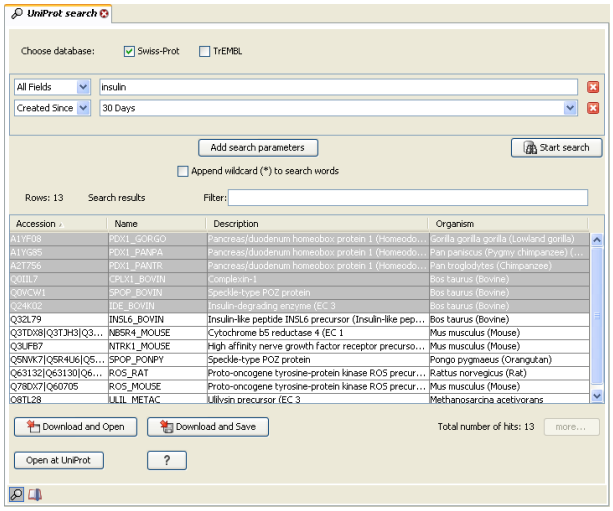


Figure 11.3: *The UniProt search view.*

11.2.1 UniProt search options

Conducting a search in **UniProt** from *CLC Protein Workbench* corresponds to conducting the search on UniProt's website. When conducting the search from *CLC Protein Workbench*, the results are available and ready to work with straight away.

Above the search fields, you can choose which database to search:

- **Swiss-Prot** This is believed to be the most accurate and best quality protein database available. All entries in the database has been currated manually and data are entered according to the original research paper.
- **TrEMBL** This database contain computer annotated protein sequences, thus the quality of the annotations is not as good as the Swiss-Prot database.

As default, *CLC Protein Workbench* offers one text field where the search parameters can be entered. Click **Add search parameters** to add more parameters to your search.
Note! The search is a "and" search, meaning that when adding search parameters to your search, you search for both (or all) text strings rather than "any" of the text strings.

You can append a wildcard character by checking the checkbox at the bottom. This means that you only have to enter the first part of the search text, e.g. searching for "genom" will find both "genomic" and "genome".

The following parameters can be added to the search:

- **All fields**. Text, searches in all parameters in the UniProt database at the same time.
- **Organism**. Text.
- **Description**. Text.
- **Created Since**. Between 30 days and 10 years.
- **Feature**. Text.

The search parameters listed in the dialog are the most recently used. The **All fields** allows searches in all parameters in the UniProt database at the same time.

When you are satisfied with the parameters you have entered, click **Start search**.

Note! When conducting a search, no files are downloaded. Instead, the program produces a list of links to the files in the UniProt database. This ensures a much faster search.

11.2.2 Handling of UniProt search results

The search result is presented as a list of links to the files in the UniProt database. The **View** displays 50 hits at a time (can be changed in the **Preferences** (see chapter 5). More hits can be displayed by clicking the **More...** button at the bottom right of the **View**. More hits can be displayed by clicking the **More...** button at the bottom left of the **View**.

Each sequence hit is represented by text in three columns:

- Accession
- Name
- Description
- Organism
- Length.

It is possible to exclude one or more of these columns by adjust the View preferences for the database search view. Furthermore, your changes in the View preferences can be saved. See section 5.5.

Several sequences can be selected, and by clicking the buttons in the bottom of the search view, you can do the following:

- Download and open, does not save the sequence.
- Download and save, lets you choose location for saving sequence.
- Open at UniProt, searches the sequence at UniProt's web page.

Double-clicking a hit will download and open the sequence. The hits can also be copied into the **View Area** or the **Navigation Area** from the search results by drag and drop, copy/paste or by using the right-click menu as described below.

Drag and drop from UniProt search results

The sequences from the search results can be opened by dragging them into a position in the **View Area**.

Note! A sequence is not saved until the **View** displaying the sequence is closed. When that happens, a dialog opens: Save changes of sequence x? (Yes or No).

The sequence can also be saved by dragging it into the **Navigation Area**. It is possible to select more sequences and drag all of them into the **Navigation Area** at the same time.

Download UniProt search results using right-click menu

You may also select one or more sequences from the list and download using the right-click menu (see figure 11.2). Choosing **Download and Save** lets you select a folder or location where the sequences are saved when they are downloaded. Choosing **Download and Open** opens a new view for each of the selected sequences.

Copy/paste from UniProt search results

When using copy/paste to bring the search results into the **Navigation Area**, the actual files are downloaded from UniProt.

To copy/paste files into the **Navigation Area**:

> **select one or more of the search results** | **Ctrl + C (⌘ + C on Mac)** | **select location or folder in the Navigation Area** | **Ctrl + V**

Note! Search results are downloaded before they are saved. Downloading and saving several files may take some time. However, since the process runs in the background (displayed in the **Toolbox** under the **Processes** tab) it is possible to continue other tasks in the program. Like the search process, the download process can be stopped, paused, and resumed.

11.2.3 Save UniProt search parameters

The search view can be saved either using dragging the search tab and and dropping it in the **Navigation Area** or by clicking **Save** (). When saving the search, only the parameters are saved - not the results of the search. This is useful if you have a special search that you perform from time to time.

Even if you don't save the search, the next time you open the search view, it will remember the parameters from the last time you did a search.

11.3 Search for structures at NCBI

This section describes searches for three dimensional structures from the NCBI structure database `http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml`. For manipulating and visualization of the downloaded structures see section 13.

The NCBI search view is opened in this way:

> **Search** | **Search for structures at NCBI** ()

or **Ctrl + B (⌘ + B on Mac)**

This opens the view shown in figure 11.4:

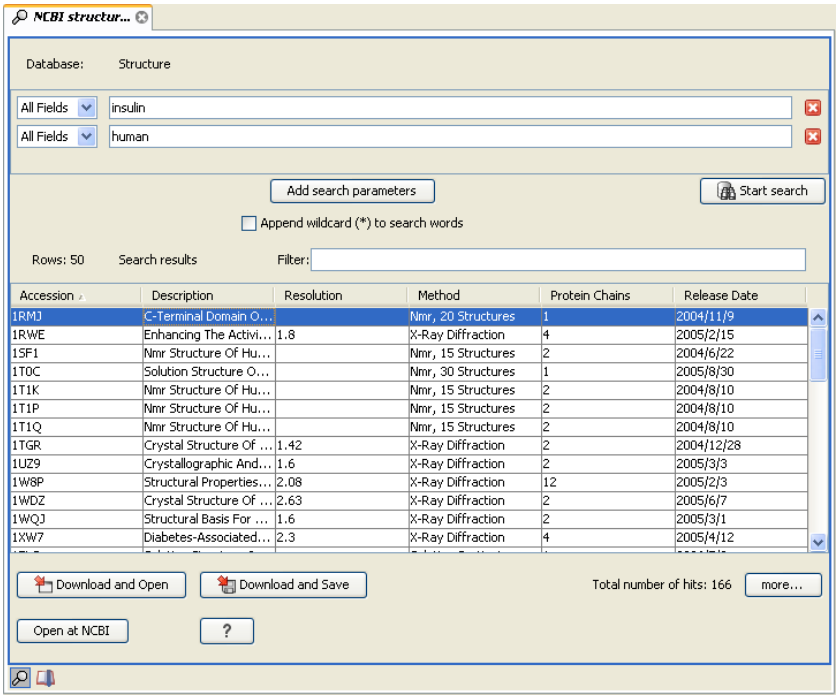

NCBI structur...
Database: Structure
All Fields
insulin
All Fields
human
Add search parameters
Start search
Append wildcard (*) to search words
Rows: 50
Search results
Filter:
Accession
Description
Resolution
Method
Protein Chains
Release Date
1RMJ C-Terminal Domain O... Nmr, 20 Structures 1 2004/11/9
1RWE Enhancing The Activi... 1.8 X-Ray Diffraction 4 2005/2/15
1SF1 Nmr Structure Of Hu... Nmr, 15 Structures 2 2004/6/22
1T0C Solution Structure O... Nmr, 30 Structures 1 2005/8/30
1T1K Nmr Structure Of Hu... Nmr, 15 Structures 2 2004/8/10
1T1P Nmr Structure Of Hu... Nmr, 15 Structures 2 2004/8/10
1T1Q Nmr Structure Of Hu... Nmr, 15 Structures 2 2004/8/10
1TGR Crystal Structure Of ... 1.42 X-Ray Diffraction 2 2004/12/28
1UZ9 Crystallographic And... 1.6 X-Ray Diffraction 2 2005/3/3
1W8P Structural Properties... 2.08 X-Ray Diffraction 12 2005/2/3
1WDZ Crystal Structure Of ... 2.63 X-Ray Diffraction 2 2005/6/7
1WQJ Structural Basis For ... 1.6 X-Ray Diffraction 2 2005/3/1
1XW7 Diabetes-Associated... 2.3 X-Ray Diffraction 4 2005/4/12
Download and Open
Download and Save
Total number of hits: 166
more...
Open at NCBI
?

Figure 11.4: *The structure search view.*

11.3.1 Structure search options

Conducting a search in the **NCBI Database** from *CLC Protein Workbench* corresponds to conducting search for structures on the NCBI's Entrez website. When conducting the search from *CLC Protein Workbench*, the results are available and ready to work with straight away.

As default, *CLC Protein Workbench* offers one text field where the search parameters can be entered. Click **Add search parameters** to add more parameters to your search.

Note! The search is a "AND" search, meaning that when adding search parameters to your search, you search for both (or all) text strings rather than "any" of the text strings.

You can append a wildcard character by clicking the checkbox at the bottom. This means that you only have to enter the first part of the search text, e.g. searching for "prot" will find both "protein" and "protease".

The following parameters can be added to the search:

- **All fields**. Text, searches in all parameters in the NCBI structure database at the same time.
- **Organism**. Text.
- **Author**. Text.
- **PdbAcc**. The accession number of the structure in the PDB database.

The search parameters are the most recently used. The **All fields** allows searches in all parameters in the database at the same time.

All fields also provide an opportunity to restrict a search to parameters which are not listed in the dialog. E.g. writing 'gene[Feature key] AND mouse' in **All fields** generates hits in the GenBank database which contains one or more genes and where 'mouse' appears somewhere in GenBank file. NB: the 'Feature Key' option is only available in GenBank when searching for nucleotide structures. For more information about how to use this syntax, see http://www.ncbi.nlm.nih.gov/entrez/query/static/help/Summary_Matrices.html#Search_Fields_and_Qualifiers

When you are satisfied with the parameters you have entered click **Start search**.

Note! When conducting a search, no files are downloaded. Instead, the program produces a list of links to the files in the NCBI database. This ensures a much faster search.

11.3.2 Handling of NCBI structure search results

The search result is presented as a list of links to the files in the NCBI database. The **View** displays 50 hits at a time (can be changed in the **Preferences** (see chapter 5). More hits can be displayed by clicking the **More...** button at the bottom right of the **View**.

Each structure hit is represented by text in three columns:

- Accession.
- Description.
- Resolution.
- Method.
- Protein chains
- Release date.

It is possible to exclude one or more of these columns by adjust the View preferences for the database search view. Furthermore, your changes in the View preferences can be saved. See section 5.5.

Several structures can be selected, and by clicking the buttons in the bottom of the search view, you can do the following:

- **Download and open.** Download and open immediately.
- **Download and save.** Download and save lets you choose location for saving structure.
- **Open at NCBI.** Open additional information on the selected structure at NCBI's web page.

Double-clicking a hit will download and open the structure. The hits can also be copied into the **View Area** or the **Navigation Area** from the search results by drag and drop, copy/paste or by using the right-click menu as described below.

Drag and drop from structure search results

The structures from the search results can be opened by dragging them into a position in the **View Area**.

Note! A structure is not saved until the **View** displaying the structure is closed. When that happens, a dialog opens: Save changes of structure x? (Yes or No).

The structure can also be saved by dragging it into the **Navigation Area**. It is possible to select more structures and drag all of them into the **Navigation Area** at the same time.

Download structure search results using right-click menu

You may also select one or more structures from the list and download using the right-click menu (see figure 11.5). Choosing **Download and Save** lets you select a folder or location where the structures are saved when they are downloaded. Choosing **Download and Open** opens a new view for each of the selected structures.

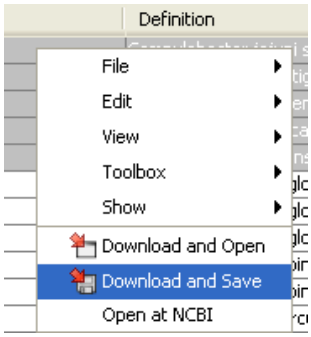


Figure 11.5: *By right-clicking a search result, it is possible to choose how to handle the relevant structure.*

The selected structures are not downloaded from the NCBI website but is downloaded from the RCSB Protein Data Bank `http://www.rcsb.org/pdb/home/home.do` in mmCIF format.

Copy/paste from structure search results

When using copy/paste to bring the search results into the **Navigation Area**, the actual files are downloaded.

To copy/paste files into the **Navigation Area**:

> **select one or more of the search results | Ctrl + C (⌘ + C on Mac) | select location or folder in the Navigation Area | Ctrl + V**

Note! Search results are downloaded before they are saved. Downloading and saving several files may take some time. However, since the process runs in the background (displayed in the **Status bar**) it is possible to continue other tasks in the program. Like the search process, the download process can be stopped. This is done in the **Toolbox** in the **Processes** tab.

11.3.3 Save structure search parameters

The search view can be saved either using dragging the search tab and and dropping it in the **Navigation Area** or by clicking **Save** (). When saving the search, only the parameters are saved

- not the results of the search. This is useful if you have a special search that you perform from time to time.

Even if you don't save the search, the next time you open the search view, it will remember the parameters from the last time you did a search.

11.4 Sequence web info

CLC Protein Workbench provides direct access to web-based search in various databases and on the Internet using your computer's default browser. You can look up a sequence in the databases of NCBI and UniProt, search for a sequence on the Internet using Google and search for Pubmed references at NCBI. This is useful for quickly obtaining updated and additional information about a sequence.

The functionality of these search functions depends on the information that the sequence contains. You can see this information by viewing the sequence as text (see section 10.5). In the following sections, we will explain this in further detail.

The procedure for searching is identical for all four search options (see also figure 11.6):

Open a sequence or a sequence list | Right-click the name of the sequence | Web Info ( ) | select the desired search function

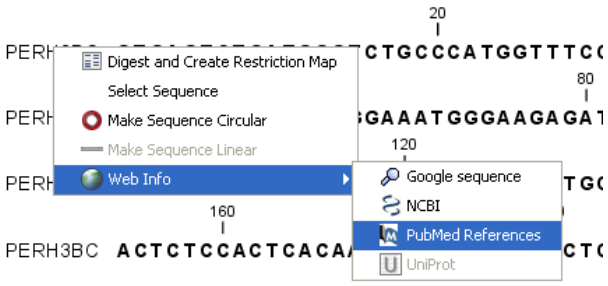


Figure 11.6: *Open webpages with information about this sequence.*

This will open your computer's default browser searching for the sequence that you selected.

11.4.1 Google sequence

The Google search function uses the accession number of the sequence which is used as search term on http://www.google.com. The resulting web page is equivalent to typing the accession number of the sequence into the search field on http://www.google.com.

11.4.2 NCBI

The NCBI search function searches in GenBank at NCBI (http://www.ncbi.nlm.nih.gov) using an identification number (when you view the sequence as text it is the "GI" number). Therefore, the sequence file must contain this number in order to look it up at NCBI. All sequences downloaded from NCBI have this number.

11.4.3 PubMed References

The PubMed references search option lets you look up Pubmed articles based on references contained in the sequence file (when you view the sequence as text it contains a number of "PUBMED" lines). Not all sequence have these PubMed references, but in this case you will se a dialog and the browser will not open.

11.4.4 UniProt

The UniProt search function searches in the UniProt database (http://www.ebi.uniprot.org) using the accession number. Furthermore, it checks whether the sequence was indeed downloaded from UniProt.

11.4.5 Additional annotation information

When sequences are downloaded from GenBank they often link to additional information on taxonomy, conserved domains etc. If such information is available for a sequence it is possible to access additional accurate online information. If the db_xref identifier line is found as part of the annotation information in the downloaded GenBank file, it is possible to easily look up additional information on the NCBI web-site.

To access this feature, simply right click an annotation and see which databases are available.

Chapter 12

BLAST Search

Contents

12.1 Running BLAST searches . . . 161
12.1.1 BLAST at NCBI . . . 162
12.1.2 BLAST a partial sequence against NCBI . . . 165
12.1.3 BLAST against local data . . . 165
12.1.4 BLAST a partial sequence against a local database . . . 167
12.2 Output from BLAST searches . . . 167
12.2.1 Graphical overview for each query sequence . . . 167
12.2.2 Overview BLAST table . . . 167
12.2.3 BLAST graphics . . . 169
12.2.4 BLAST table . . . 170
12.3 Local BLAST databases . . . 172
12.3.1 Make pre-formatted BLAST databases available . . . 173
12.3.2 Download NCBI pre-formatted BLAST databases . . . 173
12.3.3 Create local BLAST databases . . . 174
12.4 Manage BLAST databases . . . 175
12.4.1 Migrating from a previous version of the Workbench . . . 176
12.5 Bioinformatics explained: BLAST . . . 176
12.5.1 Examples of BLAST usage . . . 177
12.5.2 Searching for homology . . . 177
12.5.3 How does BLAST work? . . . 177
12.5.4 Which BLAST program should I use? . . . 179
12.5.5 Which BLAST options should I change? . . . 180
12.5.6 Explanation of the BLAST output . . . 181
12.5.7 I want to BLAST against my own sequence database, is this possible? . . . 183
12.5.8 What you cannot get out of BLAST . . . 184
12.5.9 Other useful resources . . . 184

CLC Protein Workbench offers to conduct BLAST searches on protein and DNA sequences. In short, a BLAST search identifies homologous sequences between your input (query) query sequence and a database of sequences [McGinnis and Madden, 2004]. BLAST (Basic Local

Alignment Search Tool), identifies homologous sequences using a heuristic method which finds short matches between two sequences. After initial match BLAST attempts to start local alignments from these initial matches.

If you are interested in the bioinformatics behind BLAST, there is an easy-to-read explanation of this in section 12.5.

With *CLC Protein Workbench* there are two ways of performing BLAST searches: You can either have the BLAST process run on NCBI's BLAST servers (http://www.ncbi.nlm.nih.gov/) or perform the BLAST search on your own computer. The advantage of running the BLAST search on NCBI servers is that you have readily access to the most popular BLAST databases without having to download them to your own computer. The advantage of running BLAST on your own computer is that you can use your own sequence data, and that this can sometimes be faster and more reliable for big batch BLAST jobs

Figure 12.8 shows an example of a BLAST result in the *CLC Protein Workbench*.

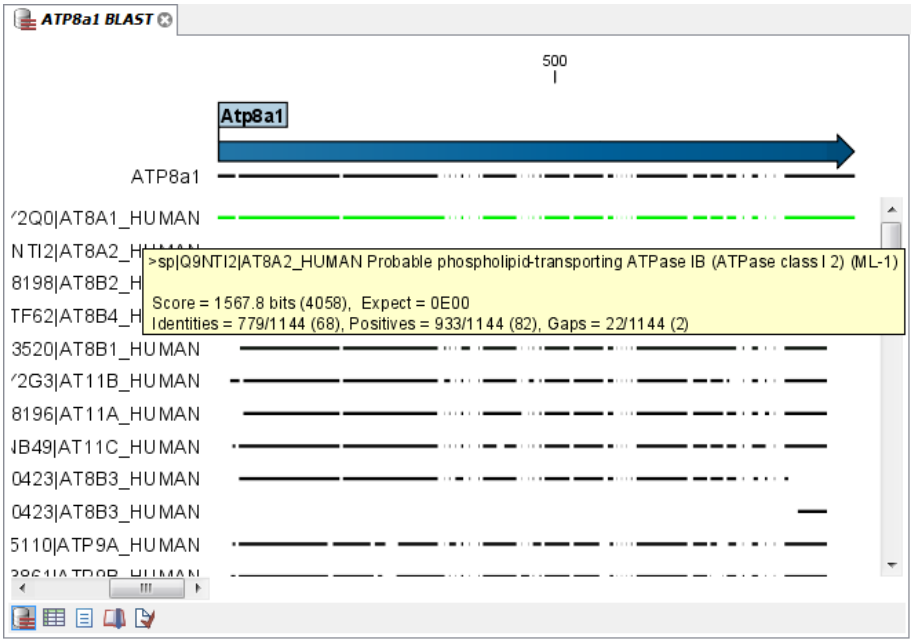


Figure 12.1: *Display of the output of a BLAST search. At the top is there a graphical representation of BLAST hits with tool-tips showing additional information on individual hits. Below is a tabular form of the BLAST results.*

12.1 Running BLAST searches

With the *CLC Protein Workbench* there are two ways of performing BLAST searches: You can either have the BLAST process run on NCBI's BLAST servers (http://www.ncbi.nlm.nih.gov/) or you can perform the BLAST search on your own computer.

The advantage of running the BLAST search on NCBI servers is that you have readily access to the popular, and often very large, BLAST databases without having to download them to your own computer. The advantages of running BLAST on your own computer include that you can use your own sequence collections as blast databases, and that running big batch BLAST jobs can be faster and more reliable when done locally.

12.1.1 BLAST at NCBI

When running a BLAST search at the NCBI, the Workbench sends the sequences you select to the NCBI's BLAST servers. When the results are ready, they will be automatically downloaded and displayed in the Workbench. When you enter a large number of sequences for searching with BLAST, the Workbench automatically splits the sequences up into smaller subsets and sends one subset at the time to NCBI. This is to avoid exceeding any internal limits the NCBI places on the number of sequences that can be submitted to them for BLAST searching. The size of the subset created in the CLC software depends both on the number and size of the sequences.

To start a BLAST job to search your sequences against databases held at the NCBI:

Toolbox | **BLAST** ( ) | **NCBI BLAST** ( )

Alternatively, use the keyboard shortcut: Ctrl+Shift+B for Windows and ⌘ +Shift+B on Mac OS.

This opens the dialog seen in figure 12.2

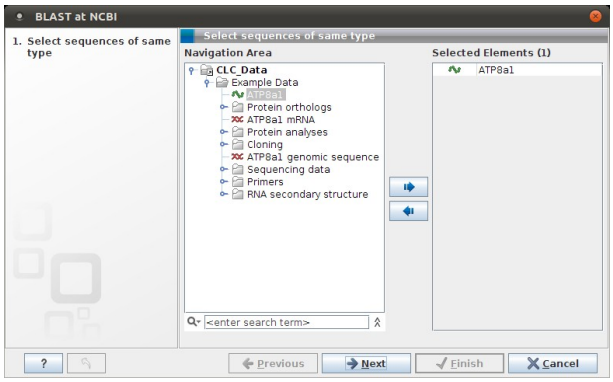


Figure 12.2: *Choose one or more sequences to conduct a BLAST search with.*

Select one or more sequences of the same type (either DNA or protein) and click **Next**.

In this dialog, you choose which type of BLAST search to conduct, and which database to search against. See figure 12.3. The databases at the NCBI listed in the dropdown box will correspond to the query sequence type you have, DNA or protein, and the type of blast search you have chosen to run. A complete list of these databases can be found in Appendix D. Here you can also read how to add additional databases available the NCBI to the list provided in the dropdown menu.

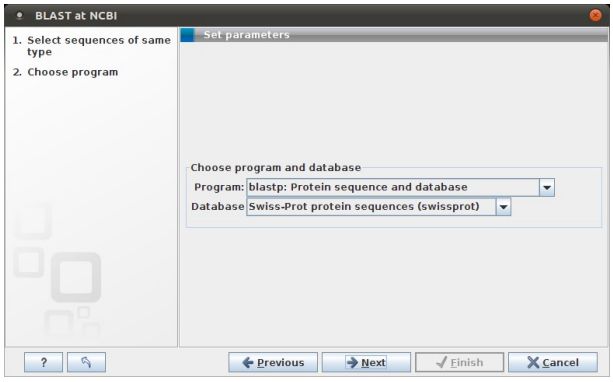


Figure 12.3: *Choose a BLAST Program and a database for the search.*

BLAST programs for DNA query sequences:

- **BLASTn: DNA sequence against a DNA database.** Used to look for DNA sequences with homologous regions to your nucleotide query sequence.
- **BLASTx: Translated DNA sequence against a Protein database.** Automatic translation of your DNA query sequence in six frames; these translated sequences are then used to search a protein database.
- **tBLASTx: Translated DNA sequence against a Translated DNA database.** Automatic translation of your DNA query sequence and the DNA database, in six frames. The resulting peptide query sequences are used to search the resulting peptide database. Note that this type of search is computationally intensive.

BLAST programs for protein query sequences:

- **BLASTp: Protein sequence against Protein database.** Used to look for peptide sequences with homologous regions to your peptide query sequence.
- **tBLASTn: Protein sequence against Translated DNA database.** Peptide query sequences are searched against an automatically translated, in six frames, DNA database.

If you search against the **Protein Data Bank protein** database homologous sequences are found to the query sequence, these can be downloaded and opened with the 3D view.

Click **Next**.

This window, see figure 12.4, allows you to choose parameters to tune your BLAST search, to meet your requirements.

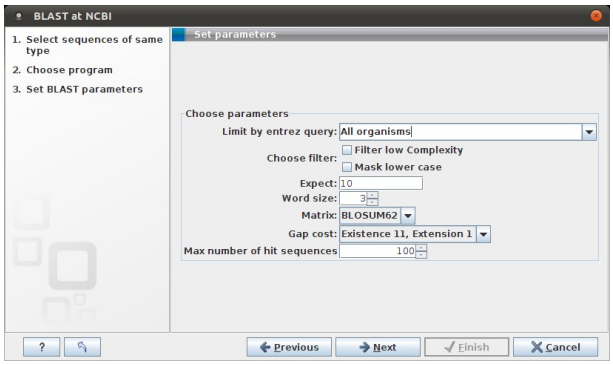


Figure 12.4: *Parameters that can be set before submitting a BLAST search.*

When choosing BLASTx or tBLASTx to conduct a search, you get the option of selecting a translation table for the genetic code. The standard genetic code is set as default. This setting is particularly useful when working with organisms or organelles that have a genetic code different from the standard genetic code.

The following description of BLAST search parameters is based on information from `http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml`.

- **Limit by Entrez query** BLAST searches can be limited to the results of an Entrez query against the database chosen. This can be used to limit searches to subsets of entries in the BLAST databases. Any terms can be entered that would normally be allowed in an Entrez search

session. More information about Entrez queries can be found at http://www.ncbi.nlm.nih.gov/books/NBK3837/#EntrezHelp.Entrez_Searching_Options. The syntax described there is the same as would be accepted in the CLC interface. Some commonly used Entrez queries are pre-entered and can be chosen in the drop down menu.

- **Choose filter**
  - **Low-complexity.** Mask off segments of the query sequence that have low compositional complexity. Filtering can eliminate statistically significant, but biologically uninteresting reports from the BLAST output (e.g. hits against common acidic-, basic- or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.
  - **Mask lower case.** If you have a sequence with regions denoted in lower case, and other regions in upper case, then choosing this option would keep any of the regions in lower case from being considered in your BLAST search.
- **Expect.** The threshold for reporting matches against database sequences: the default value is 10, meaning that under the circumstances of this search, 10 matches are expected to be found merely by chance according to the stochastic model of Karlin and Altschul (1990). Details of how E-values are calculated can be found at the NCBI: http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html If the E-value ascribed to a match is greater than the EXPECT threshold, the match will not be reported. Lower EXPECT thresholds are more stringent, leading to fewer chance matches being reported. Increasing the threshold results in more matches being reported, but many may just matching by chance, not due to any biological similarity. Values of E less than one can be entered as decimals, or in scientific notiation. For example, 0.001, 1e-3 and 10e-4 would be equivalent and acceptable values.
- **Word Size.** BLAST is a heuristic that works by finding word-matches between the query and database sequences. You may think of this process as finding "hot-spots" that BLAST can then use to initiate extensions that might lead to full-blown alignments. For nucleotide-nucleotide searches (i.e. "BLASTn") an exact match of the entire word is required before an extension is initiated, so that you normally regulate the sensitivity and speed of the search by increasing or decreasing the wordsize. For other BLAST searches non-exact word matches are taken into account based upon the similarity between words. The amount of similarity can be varied so that you normally uses just the wordsizes 2 and 3 for these searches.
- **Matrix.** A key element in evaluating the quality of a pairwise sequence alignment is the "substitution matrix", which assigns a score for aligning any possible pair of residues. The matrix used in a BLAST search can be changed depending on the type of sequences you are searching with (see the BLAST Frequently Asked Questions). Only applicable for protein sequences or translated DNA sequences.
- **Gap Cost.** The pull down menu shows the Gap Costs (Penalty to open Gap and penalty to extend Gap). Increasing the Gap Costs and Lambda ratio will result in alignments which decrease the number of Gaps introduced.
- **Max number of hit sequences.** The maximum number of database sequences, where BLAST found matches to your query sequence, to be included in the BLAST report.

The parameters you choose will affect how long BLAST takes to run. A search of a small database, requesting only hits that meet stringent criteria will generally be quite quick. Searching large databases, or allowing for very remote matches, will of course take longer.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

12.1.2 BLAST a partial sequence against NCBI

You can search a database using only a part of a sequence directly from the sequence view:

select the sequence region to send to BLAST | **right-click the selection** | **BLAST Selection Against NCBI (****)**

This will go directly to the dialog shown in figure 12.3 and the rest of the options are the same as when performing a BLAST search with a full sequence.

12.1.3 BLAST against local data

Running BLAST searches on your local machine can have several advantages over running the searches remotely at the NCBI:

- It can be faster.
- It does not rely on having a stable internet connection.
- It does not depend on the availability of the NCBI BLAST blast servers.
- You can use longer query sequences.
- You use your own data sets to search against.

On a technical level, the *CLC Protein Workbench* uses the NCBI's blast+ software (see `ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/`). Thus, the results of using a particular data set to search the same database, with the same search parameters, would give the same results, whether run locally or at the NCBI.

There are a number of options for what you can search against:

- You create a database based on data already imported into your Workbench (see section 12.3.3)
- You can add pre-formatted databases (see section 12.3.1)
- You can use sequence data from the **Navigation Area** directly, without creating a database first.

To conduct a BLAST search:

or **Toolbox** | **BLAST (** **)** | **Local BLAST (** **)**

This opens the dialog seen in figure 12.5:

Select one or more sequences of the same type (DNA or protein) and click **Next**.

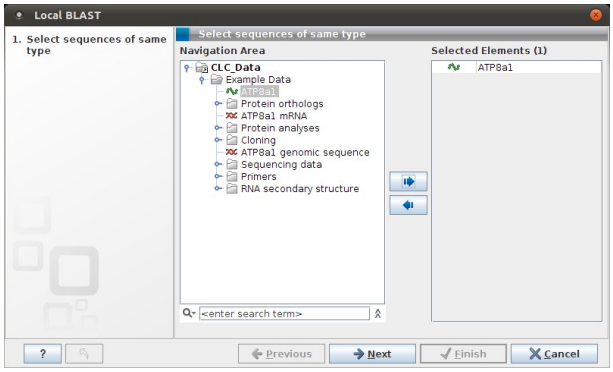


Figure 12.5: *Choose one or more sequences to conduct a BLAST search.*

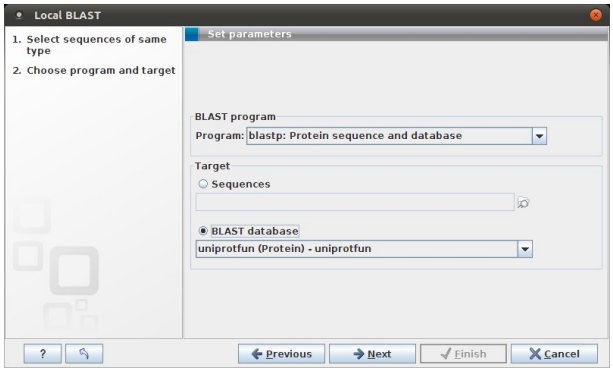


Figure 12.6: *Choose a BLAST program and a target database.*

This opens the dialog seen in figure 12.6:

At the top, you can choose between different BLAST programs. See section 12.1.1 for information about these methods.

You then specify the target database to use:

- **Sequences**. When you choose this option, you can use sequence data from the **Navigation Area** as database by clicking the **Browse and select** icon ( ). A temporary BLAST database will be created from these sequences and used for the BLAST search. It is deleted afterwards. If you want to be able to click in the BLAST result to retrieve the hit sequences from the BLAST database at a later point, you should *not* use this option; create a create a BLAST database first, see section 12.3.3.
- **BLAST Database**. Select a database already available in one of your designated BLAST database folders. Read more in section 12.4.

When a database or a set of sequences has been selected, click **Next**.

This opens the dialog seen in figure 12.7:

See section 12.1.1 for information about these limitations.

There is one setting available for local BLAST jobs that is not relevant for remote searches at the NCBI:

- **Number of processors.** You can specify the number of processors which should be used if your Workbench is installed on a multi-processor system.

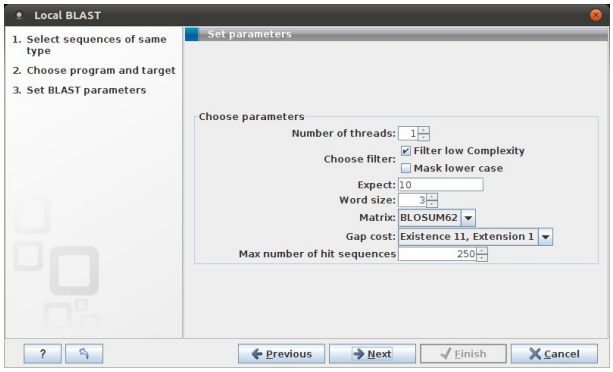


Figure 12.7: *Examples of parameters that can be set before submitting a BLAST search.*

12.1.4 BLAST a partial sequence against a local database

You can search a database using only a part of a sequence directly from the sequence view:

select the region that you wish to BLAST | right-click the selection | BLAST Selection Against Local Database ( )

This will go directly to the dialog shown in figure 12.6 and the rest of the options are the same as when performing a BLAST search with a full sequence.

12.2 Output from BLAST searches

The output of a BLAST search is similar whether you have chosen to run your search locally or at the NCBI. If a single query sequence was used, then the results will show the hits found in that database with that single sequence. If more than one sequence was used to query a database, the default view of the results is a summary table, showing the description of the top database hit against each query sequence, and the number of hits found.

12.2.1 Graphical overview for each query sequence

Double clicking on a given row of a tabular blast table opens a graphical overview of the blast results for a particular query sequence, as shown in figure figure 12.8. In cases where only one sequence was entered into a BLAST search, such a graphical overview is the default output.

Figure 12.8 shows an example of a BLAST result for an individual query sequence in the *CLC Protein Workbench*.

Detailed descriptions of the overview BLAST table and the graphical BLAST results view are described below.

12.2.2 Overview BLAST table

In the overview BLAST table for a multi-sequence blast search, as shown in figure 12.9, there is one row for each query sequence. Each row represents the BLAST result for this query sequence.

Double-clicking a row will open the BLAST result for this query sequence, allowing more detailed investigation of the result. You can also select one or more rows and click the **Open BLAST Output** button at the bottom of the view. Clicking the **Open Query Sequence** will open a sequence

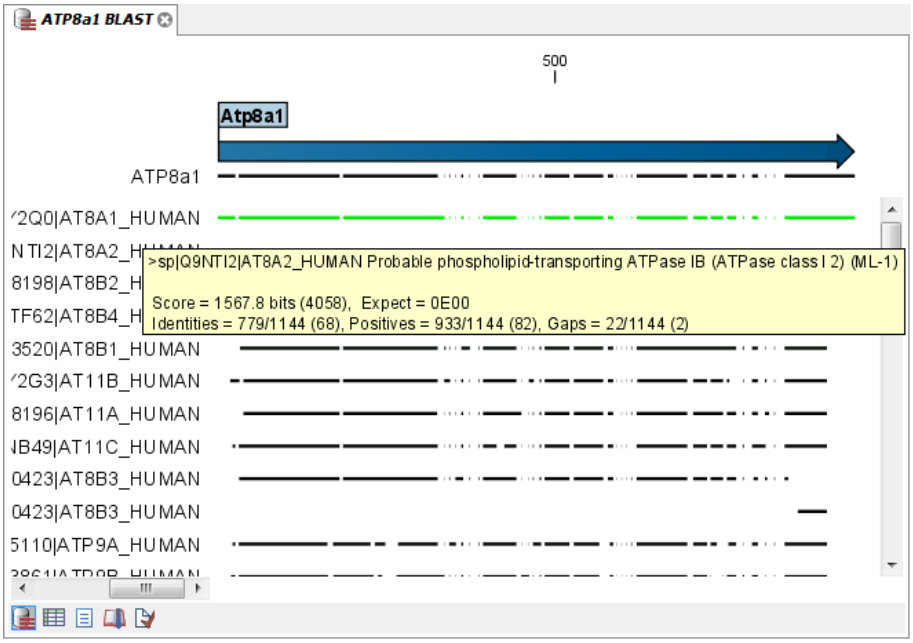


Figure 12.8: *Default display of the output of a BLAST search for one query sequence. At the top is there a graphical representation of BLAST hits with tool-tips showing additional information on individual hits.*

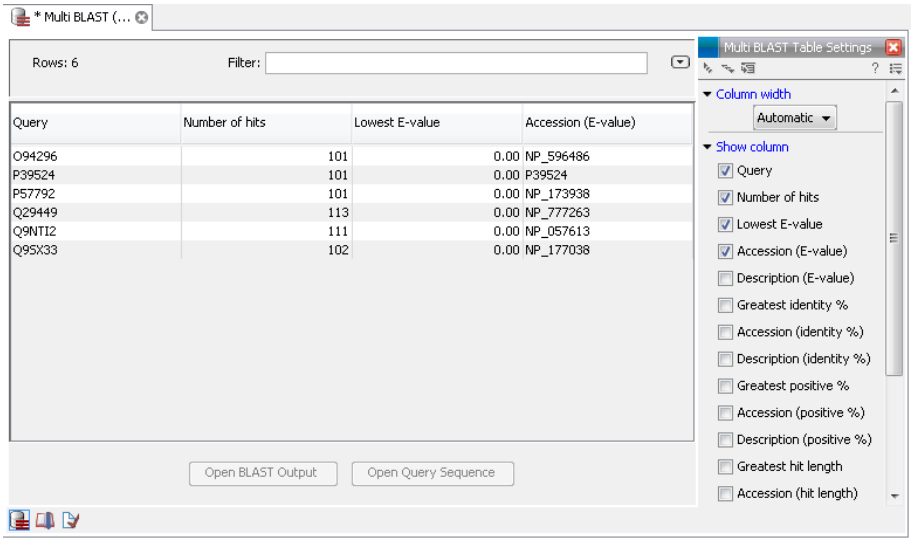


Figure 12.9: *An overview BLAST table summarizing the results for a number of query sequences.*

list with the selected query sequences. This can be useful in work flows where BLAST is used as a filtering mechanism where you can filter the table to include e.g. sequences that have a certain top hit and then extract those.

In the overview table, the following information is shown:

- Query: Since this table displays information about several query sequences, the first column is the name of the query sequence.
- Number of hits: The number of hits for this query sequence.
- For the following list, the value of the best hit is displayed together with accession number

and description of this hit.

  - Lowest E-value
  - Greatest identity %
  - Greatest positive %
  - Greatest hit length
  - Greatest bit score

If you wish to save some of the BLAST results as individual elements in the **Navigation Area**, open them and click **Save As** in the **File** menu.

12.2.3 BLAST graphics

The **BLAST editor** shows the sequences hits which were found in the BLAST search. The hit sequences are represented by colored horizontal lines, and when hovering the mouse pointer over a BLAST hit sequence, a tooltip appears, listing the characteristics of the sequence. As default, the query sequence is fitted to the window width, but it is possible to zoom in the windows and see the actual sequence alignments returned from the BLAST server.

There are several settings available in the **BLAST Graphics** view.

- **BLAST Layout.** You can choose to **Gather sequences at top**. Enabling this option affects the view that is shown when scrolling horizontally along a BLAST result. If selected, the sequence hits which did not contribute to the visible part of the BLAST graphics will be omitted whereas the found BLAST hits will automatically be placed right below the query sequence.
- **Compactness**: You can control the level of sequence detail to be displayed:
  - **Not compact.** Full detail and spaces between the sequences.
  - **Low.** The normal settings where the residues are visible (when zoomed in) but with no extra spaces between.
  - **Medium.** The sequences are represented as lines and the residues are not visible. There is some space between the sequences.
  - **Compact.** Even less space between the sequences.
- **BLAST hit coloring.** You can choose whether to color hit sequences and you can adjust the coloring.
- **Coverage**: In the Alignment info in the Side Panel, you can visualize the number of hit sequences at a given position on the query sequence. The level of coverage is relative to the overall number of hits included in the result.
  - **Foreground color.** Colors the letters using a gradient, where the left side color is used for low coverage and the right side is used for maximum coverage.
  - **Background color.** Colors the background of the letters using a gradient, where the left side color is used for low coverage and the right side is used for maximum coverage
  - **Graph.** The coverage is displayed as a graph beneath the query sequence (Learn how to export the data behind the graph in section 7.4).

  - **Height.** Specifies the height of the graph.
  - **Type.** The graph can be displayed as Line plot, Bar plot or as a Color bar.
  - **Color box.** For Line and Bar plots, the color of the plot can be set by clicking the color box. If a Color bar is chosen, the color box is replaced by a gradient color box as described under Foreground color.

The remaining View preferences for BLAST Graphics are the same as those of alignments. See section 18.2.

Some of the information available in the tooltips is:

- **Name of sequence.** Here is shown some additional information of the sequence which was found. This line corresponds to the description line in GenBank (if the search was conducted on the nr database).
- **Score.** This shows the bit score of the local alignment generated through the BLAST search.
- **Expect.** Also known as the E-value. A low value indicates a homologous sequence. Higher E-values indicate that BLAST found a less homologous sequence.
- **Identities.** This number shows the number of identical residues or nucleotides in the obtained alignment.
- **Gaps.** This number shows whether the alignment has gaps or not.
- **Strand.** This is only valid for nucleotide sequences and show the direction of the aligned strands. Minus indicate a complementary strand.
- **Query.** This is the sequence (or part of the sequence) which you have used for the BLAST search.
- **Sbjct (subject).** This is the sequence found in the database.

The numbers of the query and subject sequences refer to the sequence positions in the submitted and found sequences. If the subject sequence has number 59 in front of the sequence, this means that 58 residues are found upstream of this position, but these are not included in the alignment.

By right clicking the sequence name in the Graphical BLAST output it is possible to download the full hits sequence from NCBI with accompanying annotations and information. It is also possible to just open the actual hit sequence in a new view.

12.2.4 BLAST table

In addition to the graphical display of a BLAST result, it is possible to view the BLAST results in a tabular view. In the tabular view, one can get a quick and fast overview of the results. Here you can also select multiple sequences and download or open all of these in one single step. Moreover, there is a link from each sequence to the sequence at NCBI. These possibilities are either available through a right-click with the mouse or by using the buttons below the table.

If the **BLAST table** view was not selected in **Step 4** of the BLAST search, the table can be shown in the following way:

Click the Show BLAST Table button ( ) at the bottom of the view

Figure 12.10 is an example of a BLAST Table.

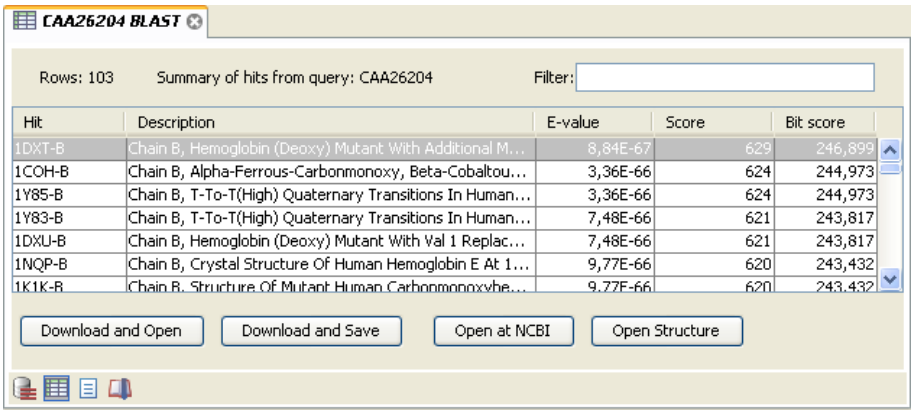


Figure 12.10: *Display of the output of a BLAST search in the tabular view. The hits can be sorted by the different columns, simply by clicking the column heading.*

The BLAST Table includes the following information:

- **Query sequence.** The sequence which was used for the search.
- **Hit.** The Name of the sequences found in the BLAST search.
- **Id.** GenBank ID.
- **Description.** Text from NCBI describing the sequence.
- **E-value.** Measure of quality of the match. Higher E-values indicate that BLAST found a less homologous sequence.
- **Score.** This shows the score of the local alignment generated through the BLAST search.
- **Bit score.** This shows the bit score of the local alignment generated through the BLAST search. Bit scores are normalized, which means that the bit scores from different alignments can be compared, even if different scoring matrices have been used.
- **Hit start.** Shows the start position in the hit sequence
- **Hit end.** Shows the end position in the hit sequence.
- **Hit length.** The length of the hit.
- **Query start.** Shows the start position in the query sequence.
- **Query end.** Shows the end position in the query sequence.
- **Overlap.** Display a percentage value for the overlap of the query sequence and hit sequence. Only the length of the local alignment is taken into account and not the full length query sequence.
- **Identity.** Shows the number of identical residues in the query and hit sequence.
- **%Identity.** Shows the percentage of identical residues in the query and hit sequence.

- **Positive.** Shows the number of similar but not necessarily identical residues in the query and hit sequence.
- **%Positive.** Shows the percentage of similar but not necessarily identical residues in the query and hit sequence.
- **Gaps.** Shows the number of gaps in the query and hit sequence.
- **%Gaps.** Shows the percentage of gaps in the query and hit sequence.
- **Query Frame/Strand.** Shows the frame or strand of the query sequence.
- **Hit Frame/Strand.** Shows the frame or strand of the hit sequence.

In the **BLAST table** view you can handle the hit sequences. Select one or more sequences from the table, and apply one of the following functions.

- **Download and Open.** Download the full sequence from NCBI and opens it. If multiple sequences are selected, they will all open (if the same sequence is listed several times, only one copy of the sequence is downloaded and opened).
- **Download and Save.** Download the full sequence from NCBI and save it. When you click the button, there will be a save dialog letting you specify a folder to save the sequences. If multiple sequences are selected, they will all open (if the same sequence is listed several times, only one copy of the sequence is downloaded and opened).
- **Open at NCBI.** Opens the corresponding sequence(s) at GenBank at NCBI. Here is stored additional information regarding the selected sequence(s). The default Internet browser is used for this purpose.
- **Open structure.** If the hit sequence contain structure information, the sequence is opened in a text view or a 3D view (3D view in CLC Protein Workbench and CLC Main Workbench).

You can do a text-based search in the information in the BLAST table by using the filter at the upper right part of the view. In this way you can search for e.g. species or other information which is typically included in the "Description" field.

The table is integrated with the graphical view described in section 12.2.3 so that selecting a hit in the table will make a selection on the corresponding sequence in the graphical view.

12.3 Local BLAST databases

BLAST databases on your local system can be made available for searches via your *CLC Protein Workbench*, (section 12.3.1). To make adding databases even easier, you can download pre-formatted BLAST databases from the NCBI from within your *CLC Protein Workbench*, (section 12.3.2). You can also easily create your own local blast databases from sequences within your *CLC Protein Workbench*, (section 12.3.3).

12.3.1 Make pre-formatted BLAST databases available

To use databases that have been downloaded or created outside the Workbench, you can either

- Put the database files in one of the locations defined in the BLAST database manager (see section 12.4)
- Add the location where your BLAST databases are stored using the BLAST database manager (see section 12.4). See figure 12.14.

12.3.2 Download NCBI pre-formatted BLAST databases

Many popular pre-formatted databases are available for download from the NCBI. You can download any of the databases available from the list at `ftp://ftp.ncbi.nih.gov/blast/db/` from within your *CLC Protein Workbench*.

You must be connected to the internet to use this tool.

If you choose:

or **Toolbox** | **BLAST** ( ) | **Download BLAST Databases** ( )

a window like the one in figure 12.11 pops up showing you the list of databases available for download.

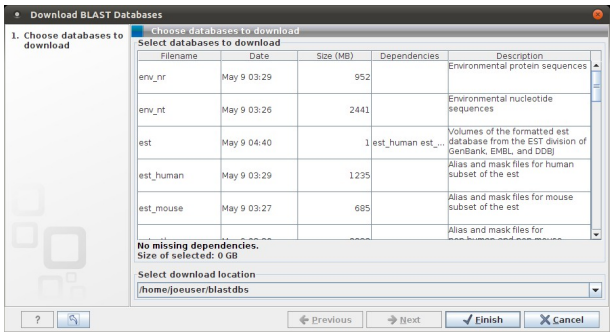


Figure 12.11: *Choose from pre-formatted BLAST databases at the NCBI available for download.*

In this window, you can see the names of the databases, the date they were made available for download on the NCBI site, the size of the files associated with that database, and a brief description of each database. You can also see whether the database has any dependencies. This aspect is described below.

You can also specify which of your database locations you would like to store the files in. Please see the **Manage BLAST Databases** section for more on this (section 12.4).

There are two very important things to note if you wish to take advantage of this tool.

- Many of the databases listed are very large. Please make sure you have room for them. If you are working on a shared system, we recommend you discuss your plans with your system administrator and fellow users.
- Some of the databases listed are dependent on others. This will be listed in the **Dependencies** column of the **Download BLAST Databases** window. This means that while

the database your are interested in may seem very small, it may require that you also download a very big database on which it depends.

An example of the second item above is *Swissprot*. To download a database from the NCBI that would allow you to search just Swissprot entries, you need to download the whole *nr* database in addition to the entry for Swissprot.

12.3.3 Create local BLAST databases

In the *CLC Protein Workbench* you can create a local database that you can use for local BLAST searches. You can specify a location on your computer to save the BLAST database files to. The Workbench will list the BLAST databases found in these locations when you set up a local BLAST search (see section 12.1.3).

DNA, RNA, and protein sequences located in the **Navigation Area** can be used to create BLAST databases from. Any given BLAST database can only include one molecule type. If you wish to use a pre-formatted BLAST database instead, see section 12.3.1.

To create a BLAST database, go to:

Toolbox | **BLAST** ( ) | **Create BLAST Database** ( )

This opens the dialog seen in figure 12.12.

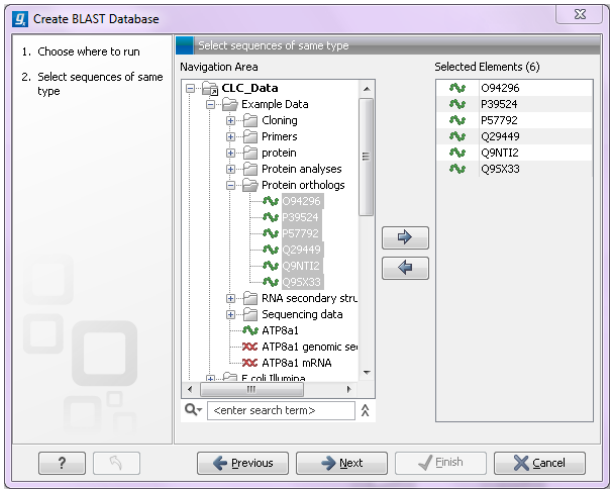


Figure 12.12: *Add sequences for the BLAST database.*

Select sequences or sequence lists you wish to include in your database and click **Next**.

In the next dialog, shown in figure 12.13, you provide the following information:

- **Name.** The name of the BLAST database. This name will be used when running BLAST searches and also as the base file name for the BLAST database files.
- **Description.** You can add more details to describe the contents of the database.
- **Location.** You can select the location to save the BLAST database files to. You can add or change the locations in this list using the **Manage BLAST Databases** tool, see section 12.4.

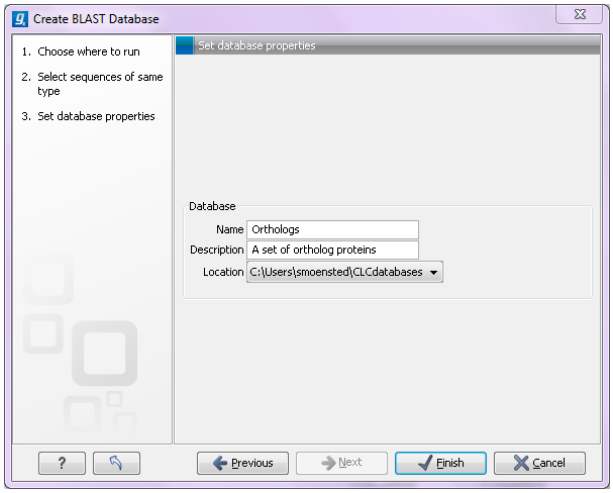


Figure 12.13: *Providing a name and description for the database, and the location to save the files to.*

Click **Finish** to create the BLAST database. Once the process is complete, the new database will be available in the **Manage BLAST Databases** dialog, see section 12.4, and when running local BLAST (see section 12.1.3).

12.4 Manage BLAST databases

The BLAST database available as targets for running local BLAST searches (see section 12.1.3) can be managed through the Manage BLAST Databases dialog (see figure 12.14):

Toolbox | **BLAST** ( ) | **Manage BLAST Databases** ( )

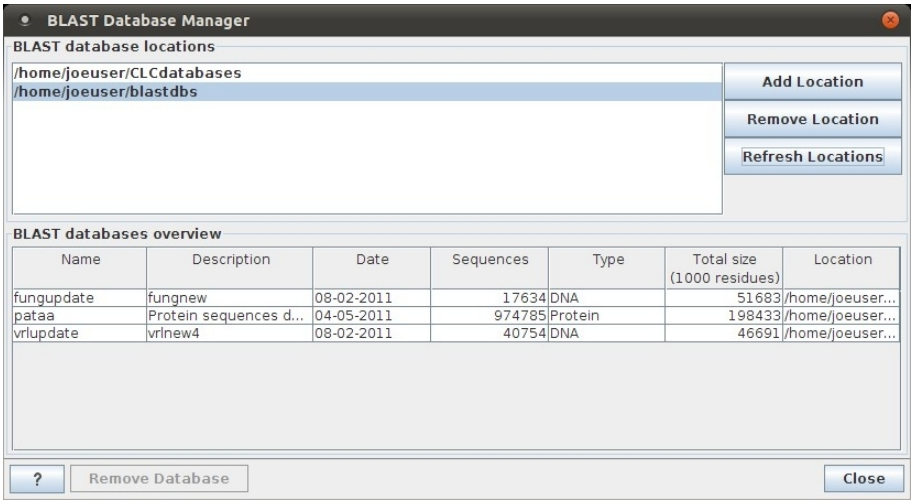


Figure 12.14: *Overview of available BLAST databases.*

At the top of the dialog, there is a list of the **BLAST database locations**. These locations are folders where the Workbench will look for valid BLAST databases. These can either be created from within the Workbench using the **Create BLAST Database tool**, see section 12.3.3, or they can be pre-formatted BLAST databases.

The list of locations can be modified using the **Add Location** and **Remove Location** buttons. Once the Workbench has scanned the locations, it will keep a cache of the databases (in order

to improve performance). If you have added new databases that are not listed, you can press **Refresh Locations** to clear the cache and search the database locations again.

By default a BLAST database location will be added under your home area in a folder called CLCdatabases. This folder is scanned recursively, through all subfolders, to look for valid databases. All other folderlocations are scanned only at the top level.

Below the list of locations, all the BLAST databases are listed with the following information:

- **Name.** The name of the BLAST database.
- **Description.** Detailed description of the contents of the database.
- **Date.** The date the database was created.
- **Sequences.** The number of sequences in the database.
- **Type.** The type can be either nucleotide (DNA) or protein.
- **Total size (1000 residues).** The number of residues in the database, either bases or amino acid.
- **Location.** The location of the database.

Below the list of BLAST databases, there is a button to **Remove Database**. This option will delete the database files belonging to the database selected.

12.4.1 Migrating from a previous version of the Workbench

In versions released before 2011, the BLAST database management was very different from this. In order to migrate from the older versions, please add the folders of the old BLAST databases as locations in the BLAST database manager (see section 12.4). The old representations of the BLAST databases in the **Navigation Area** can be deleted.

If you have saved the BLAST databases in the default folder, they will automatically appear because the default database location used in *CLC Protein Workbench 5.8* is the same as the default folder specified for saving BLAST databases in the old version.

12.5 Bioinformatics explained: BLAST

BLAST (Basic Local Alignment Search Tool) has become the *defacto* standard in search and alignment tools [Altschul et al., 1990]. The BLAST algorithm is still actively being developed and is one of the most cited papers ever written in this field of biology. Many researchers use BLAST as an initial screening of their sequence data from the laboratory and to get an idea of what they are working on. BLAST is far from being basic as the name indicates; it is a highly advanced algorithm which has become very popular due to availability, speed, and accuracy. In short, a BLAST search identifies homologous sequences by searching one or more databases usually hosted by NCBI (`http://www.ncbi.nlm.nih.gov/`), on the query sequence of interest [McGinnis and Madden, 2004].

BLAST is an open source program and anyone can download and change the program code. This has also given rise to a number of BLAST derivatives; WU-BLAST is probably the most commonly used [Altschul and Gish, 1996].

BLAST is highly scalable and comes in a number of different computer platform configurations which makes usage on both small desktop computers and large computer clusters possible.

12.5.1 Examples of BLAST usage

BLAST can be used for a lot of different purposes. A few of them are mentioned below.

- **Looking for species.** If you are sequencing DNA from unknown species, BLAST may help identify the correct species or homologous species.
- **Looking for domains.** If you BLAST a protein sequence (or a translated nucleotide sequence) BLAST will look for known domains in the query sequence.
- **Looking at phylogeny.** You can use the BLAST web pages to generate a phylogenetic tree of the BLAST result.
- **Mapping DNA to a known chromosome.** If you are sequencing a gene from a known species but have no idea of the chromosome location, BLAST can help you. BLAST will show you the position of the query sequence in relation to the hit sequences.
- **Annotations.** BLAST can also be used to map annotations from one organism to another or look for common genes in two related species.

12.5.2 Searching for homology

Most research projects involving sequencing of either DNA or protein have a requirement for obtaining biological information of the newly sequenced and maybe unknown sequence. If the researchers have no prior information of the sequence and biological content, valuable information can often be obtained using BLAST. The BLAST algorithm will search for homologous sequences in predefined and annotated databases of the users choice.

In an easy and fast way the researcher can gain knowledge of gene or protein function and find evolutionary relations between the newly sequenced DNA and well established data.

After the BLAST search the user will receive a report specifying found homologous sequences and their local alignments to the query sequence.

12.5.3 How does BLAST work?

BLAST identifies homologous sequences using a heuristic method which initially finds short matches between two sequences; thus, the method does not take the entire sequence space into account. After initial match, BLAST attempts to start local alignments from these initial matches. This also means that BLAST does not guarantee the optimal alignment, thus some sequence hits may be missed. In order to find optimal alignments, the Smith-Waterman algorithm should be used (see below). In the following, the BLAST algorithm is described in more detail.

Seeding

When finding a match between a query sequence and a hit sequence, the starting point is the *words* that the two sequences have in common. A word is simply defined as a number of letters.

For blastp the default word size is 3 *W=3*. If a query sequence has a QWRTG, the searched words are QWR, WRT, RTG. See figure 12.15 for an illustration of words in a protein sequence.

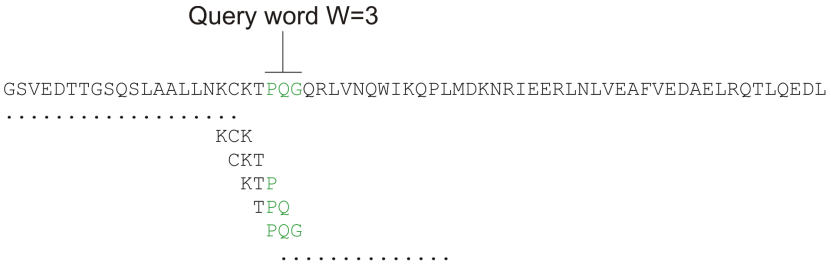


Figure 12.15: *Generation of exact BLAST words with a word size of W=3.*

During the initial BLAST seeding, the algorithm finds all common words between the query sequence and the hit sequence(s). Only regions with a word hit will be used to build on an alignment.

BLAST will start out by making words for the entire query sequence (see figure 12.15). For each word in the query sequence, a compilation of neighborhood words, which exceed the threshold of *T*, is also generated.

A neighborhood word is a word obtaining a score of at least *T* when comparing, using a selected scoring matrix (see figure 12.16). The default scoring matrix for blastp is BLOSUM62 (for explanation of scoring matrices, see www.clcbio.com/be). The compilation of exact words and neighborhood words is then used to match against the database sequences.

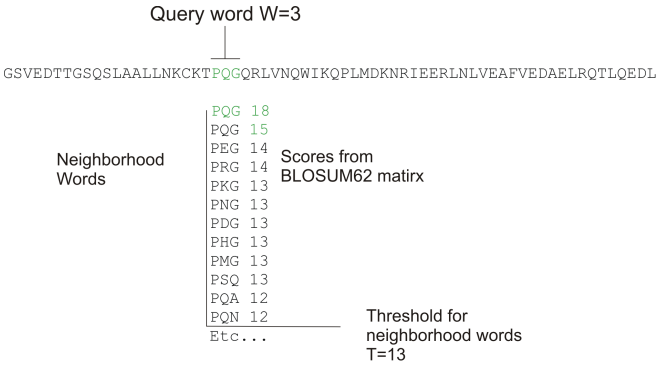
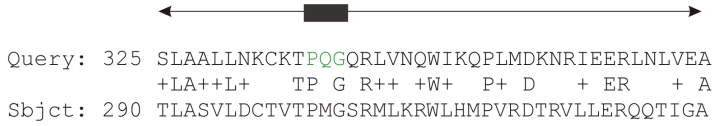


Figure 12.16: *Neighborhood BLAST words based on the BLOSUM62 matrix. Only words where the threshold* T *exceeds 13 are included in the initial seeding.*

After initial finding of words (seeding), the BLAST algorithm will extend the (only 3 residues long) alignment in both directions (see figure 12.17). Each time the alignment is extended, an alignment score is increases/decreased. When the alignment score drops below a predefined threshold, the extension of the alignment stops. This ensures that the alignment is not extended to regions where only very poor alignment between the query and hit sequence is possible. If the obtained alignment receives a score above a certain threshold, it will be included in the final BLAST result.

By tweaking the word size *W* and the neighborhood word threshold *T*, it is possible to limit the search space. E.g. by increasing *T*, the number of neighboring words will drop and thus limit the search space as shown in figure 12.18.

This will increase the speed of BLAST significantly but may result in loss of sensitivity. Increasing the word size *W* will also increase the speed but again with a loss of sensitivity.



```
Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNLVEA     365
           +LA++L+   TP G R++ +W+  P+ D   + ER   + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA     330
```

Figure 12.17: *Blast aligning in both directions. The initial word match is marked green.*

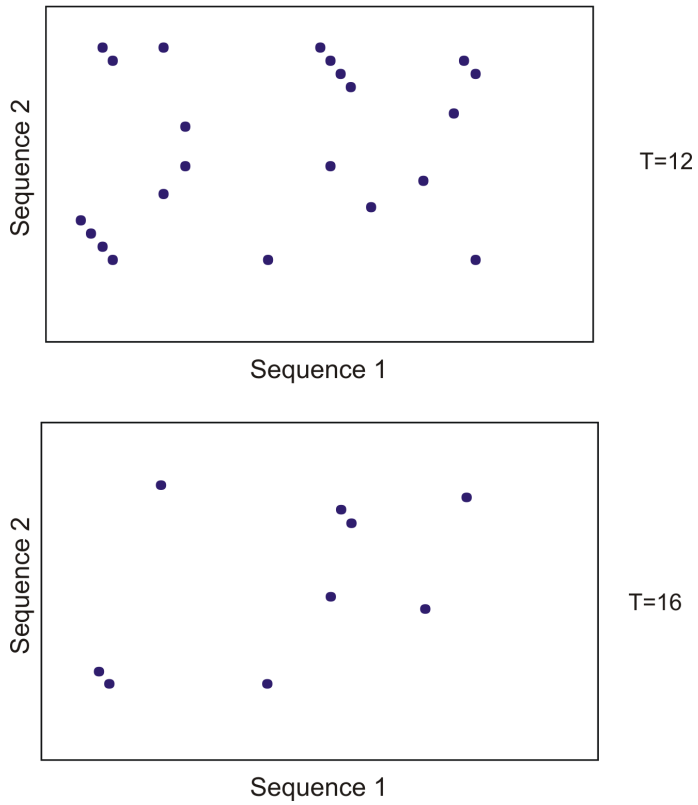


Figure 12.18: *Each dot represents a word match. Increasing the threshold of* T *limits the search space significantly.*

12.5.4 Which BLAST program should I use?

Depending on the nature of the sequence it is possible to use different BLAST programs for the database search. There are five versions of the BLAST program, blastn, blastp, blastx, tblastn, tblastx:

Option	Query Type	DB Type	Comparison	Note
blastn	Nucleotide	Nucleotide	Nucleotide-Nucleotide	
blastp	Protein	Protein	Protein-Protein	
tblastn	Protein	Nucleotide	Protein-Protein	The database is translated into protein
blastx	Nucleotide	Protein	Protein-Protein	The queries are translated into protein
tblastx	Nucleotide	Nucleotide	Protein-Protein	The queries and database are translated into protein

The most commonly used method is to BLAST a nucleotide sequence against a nucleotide database (blastn) or a protein sequence against a protein database (blastp). But often another BLAST program will produce more interesting hits. E.g. if a nucleotide sequence is translated

before the search, it is more likely to find better and more accurate hits than just a blastn search. One of the reasons for this is that protein sequences are evolutionarily more conserved than nucleotide sequences. Another good reason for translating the query sequence before the search is that you get protein hits which are likely to be annotated. Thus you can directly see the protein function of the sequenced gene.

12.5.5 Which BLAST options should I change?

The NCBI BLAST web pages and the BLAST command line tool offer a number of different options which can be changed in order to obtain the best possible result. Changing these parameters can have a great impact on the search result. It is not the scope of this document to comment on all of the options available but merely the options which can be changed with a direct impact on the search result.

The E-value

The *expect value*(E-value) can be changed in order to limit the number of hits to the most significant ones. The lower the E-value, the better the hit. The E-value is dependent on the length of the query sequence and the size of the database. For example, an alignment obtaining an E-value of 0.05 means that there is a 5 in 100 chance of occurring by chance alone.

E-values are very dependent on the query sequence length and the database size. Short identical sequence may have a high E-value and may be regarded as "false positive" hits. This is often seen if one searches for short primer regions, small domain regions etc. The default threshold for the E-value on the BLAST web page is 10. Increasing this value will most likely generate more hits. Below are some rules of thumb which can be used as a guide but should be considered with common sense.

- **E-value < 10e-100** Identical sequences. You will get long alignments across the entire query and hit sequence.
- **10e-100 < E-value < 10e-50** Almost identical sequences. A long stretch of the query protein is matched to the database.
- **10e-50 < E-value < 10e-10** Closely related sequences, could be a domain match or similar.
- **10e-10 < E-value < 1** Could be a true homologue but it is a gray area.
- **E-value > 1** Proteins are most likely not related
- **E-value > 10** Hits are most likely junk unless the query sequence is very short.

Gap costs

For blastp it is possible to specify gap cost for the chosen substitution matrix. There is only a limited number of options for these parameters. The *open gap cost* is the price of introducing gaps in the alignment, and *extension gap cost* is the price of every extension past the initial opening gap. Increasing the gap costs will result in alignments with fewer gaps.

Filters

It is possible to set different filter options before running the BLAST search. Low-complexity regions have a very simple composition compared to the rest of the sequence and may result in problems during the BLAST search [Wootton and Federhen, 1993]. A low complexity region of a protein can for example look like this 'fftfflllsss', which in this case is a region as part of a signal peptide. In the output of the BLAST search, low-complexity regions will be marked in lowercase gray characters (default setting). The low complexity region cannot be thought of as a significant match; thus, disabling the low complexity filter is likely to generate more hits to sequences which are not truly related.

Word size

Change of the word size has a great impact on the seeded sequence space as described above. But one can change the word size to find sequence matches which would otherwise not be found using the default parameters. For instance the word size can be decreased when searching for primers or short nucleotides. For blastn a suitable setting would be to decrease the default word size of 11 to 7, increase the E-value significantly (1000) and turn off the complexity filtering.

For blastp a similar approach can be used. Decrease the word size to 2, increase the E-value and use a more stringent substitution matrix, e.g. a PAM30 matrix.

Fortunately, the optimal search options for finding short, nearly exact matches can already be found on the BLAST web pages http://www.ncbi.nlm.nih.gov/BLAST/.

Substitution matrix

For protein BLAST searches, a default substitution matrix is provided. If you are looking at distantly related proteins, you should either choose a high-numbered PAM matrix or a low-numbered BLOSUM matrix. See *Bioinformatics Explained* on scoring matrices on http://www.clcbio.com/be/. The default scoring matrix for blastp is BLOSUM62.

12.5.6 Explanation of the BLAST output

The BLAST output comes in different flavors. On the NCBI web page the default output is html, and the following description will use the html output as example. Ordinary text and xml output for easy computational parsing is also available.

The default layout of the NCBI BLAST result is a graphical representation of the hits found, a table of sequence identifiers of the hits together with scoring information, and alignments of the query sequence and the hits.

The graphical output (shown in figure 12.19) gives a quick overview of the query sequence and the resulting hit sequences. The hits are colored according to the obtained alignment scores.

The table view (shown in figure 12.20) provides more detailed information on each hit and furthermore acts as a hyperlink to the corresponding sequence in GenBank.

In the alignment view one can manually inspect the individual alignments generated by the BLAST algorithm. This is particularly useful for detailed inspection of the sequence hit found(sbjct) and the corresponding alignment. In the alignment view, all scores are described for each alignment,

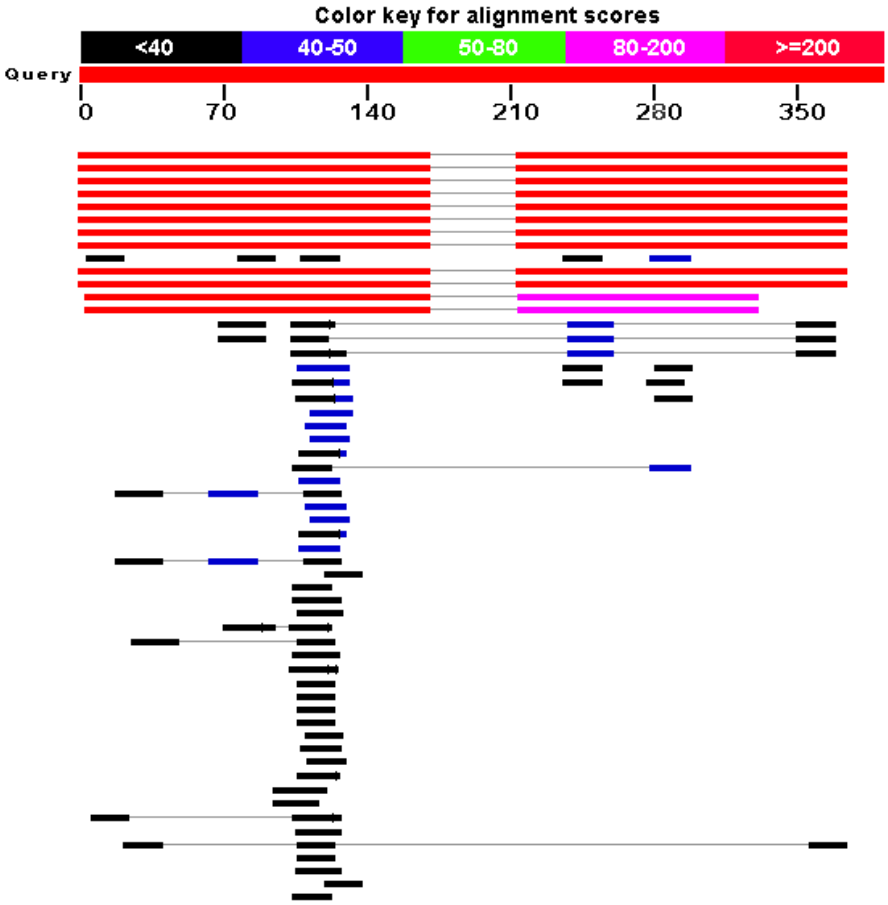


Figure 12.19: *BLAST graphical view. A simple graphical overview of the hits found aligned to the query sequence. The alignments are color coded ranging from black to red as indicated in the color label at the top.*

Sequences producing significant alignments:
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_174886.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	U E G M
NM_173210.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	U E G M
NM_173209.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	U E G M
NM_173211.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	U E G M
NM_173207.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	U E G M
NM_173208.1	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	U E G M
NM_170695.2	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	U E G M
NM_003244.2	Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF),	339	563	85%	1e-90	100%	U E G M
NM_003246.2	Homo sapiens thrombospondin 1 (THBS1), mRNA	38.2	38.2	4%	7.2	100%	U E G M
NM_177965.2	Homo sapiens chromosome 8 open reading frame 37 (C8orf37),	38.2	38.2	4%	7.2	100%	U E G M
Genomic sequences [show first]							
NT_010859.14	Homo sapiens chromosome 18 genomic contig, reference assembly	339	602	85%	1e-90	100%	
NW_926940.1	Homo sapiens chromosome 18 genomic contig, alternate assembly	339	602	85%	1e-90	100%	
NT_011109.15	Homo sapiens chromosome 19 genomic contig, reference assembly	262	375	73%	3e-67	94%	
NW_927217.1	Homo sapiens chromosome 19 genomic contig, alternate assembly	262	375	73%	3e-67	94%	

Figure 12.20: *BLAST table view. A table view with one row per hit, showing the accession number and description field from the sequence file together with BLAST output scores.*

and the start and stop positions for the query and hit sequence are listed. The strand and orientation for query sequence and hits are also found here.

In most cases, the table view of the results will be easier to interpret than tens of sequence alignments.

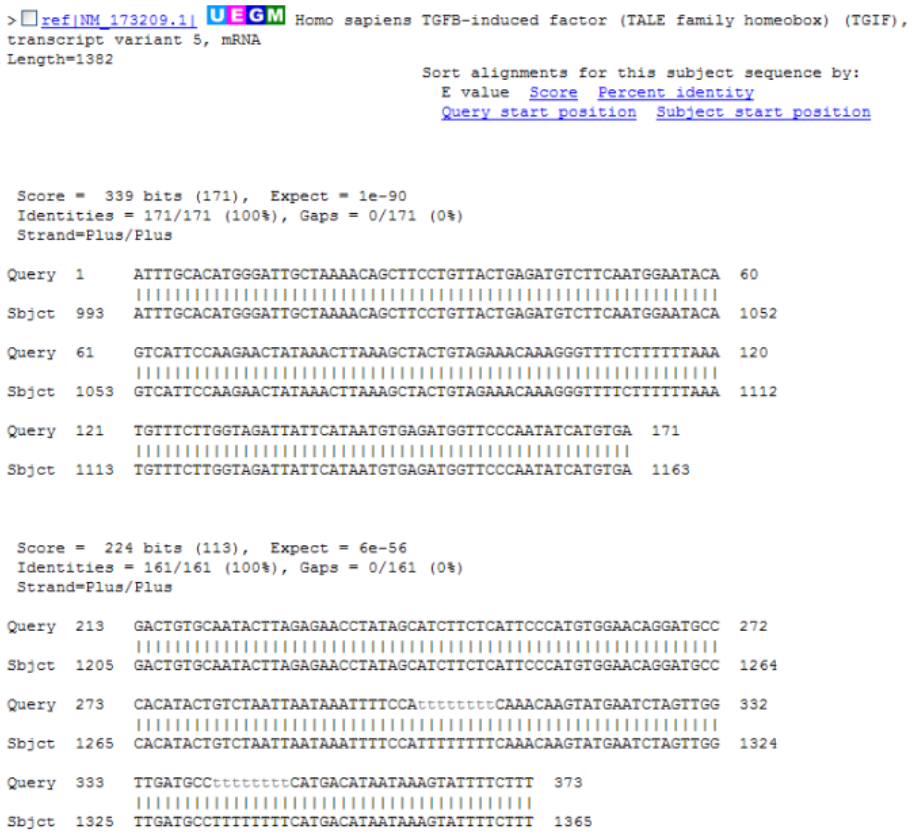

>ref|NM_173209.1| Homo sapiens TGFB-induced factor (TALE family homeobox) (TGIF), transcript variant 5, mRNA
Length=1382
Sort alignments for this subject sequence by:
E value Score Percent identity
Query start position Subject start position
Score = 339 bits (171), Expect = 1e-90
Identities = 171/171 (100%), Gaps = 0/171 (0%)
Strand=Plus/Plus
Query 1 ATTTGCACATGGGATTGCTAAAACAGCTTCCTGTTACTGAGATGTCTTCAATGGAATACA 60
Sbjct 993 ATTTGCACATGGGATTGCTAAAACAGCTTCCTGTTACTGAGATGTCTTCAATGGAATACA 1052
Query 61 GTCATTCCAAGAACTATAAACTTAAAGCTACTGTAGAAACAAAGGGTTTTCTTTTTTAAA 120
Sbjct 1053 GTCATTCCAAGAACTATAAACTTAAAGCTACTGTAGAAACAAAGGGTTTTCTTTTTTAAA 1112
Query 121 TGTTTCTTGGTAGATTATTCATAATGTGAGATGGTTCCCAATATCATGTGA 171
Sbjct 1113 TGTTTCTTGGTAGATTATTCATAATGTGAGATGGTTCCCAATATCATGTGA 1163
Score = 224 bits (113), Expect = 6e-56
Identities = 161/161 (100%), Gaps = 0/161 (0%)
Strand=Plus/Plus
Query 213 GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC 272
Sbjct 1205 GACTGTGCAATACTTAGAGAACCTATAGCATCTTCTCATTCCCATGTGGAACAGGATGCC 1264
Query 273 CACATACTGTCTAATTAATAAATTTTCCAtttttttCAAACAAGTATGAATCTAGTTGG 332
Sbjct 1265 CACATACTGTCTAATTAATAAATTTTCCATTTTTTTTCAAACAAGTATGAATCTAGTTGG 1324
Query 333 TTGATGCCtttttttCATGACATAATAAAGTATTTTCTTT 373
Sbjct 1325 TTGATGCCTTTTTTTTCATGACATAATAAAGTATTTTCTTT 1365

Figure 12.21: *Alignment view of BLAST results. Individual alignments are represented together with BLAST scores and more.*

12.5.7 I want to BLAST against my own sequence database, is this possible?

It is possible to download the entire BLAST program package and use it on your own computer, institution computer cluster or similar. This is preferred if you want to search in proprietary sequences or sequences unavailable in the public databases stored at NCBI. The downloadable BLAST package can either be installed as a web-based tool or as a command line tool. It is available for a wide range of different operating systems.

The BLAST package can be downloaded free of charge from the following location `http://www.ncbi.nlm.nih.gov/BLAST/download.shtml`

Pre-formatted databases are available from a dedicated BLAST ftp site `ftp://ftp.ncbi.nlm.nih.gov/blast/db/`. Moreover, it is possible to download programs/scripts from the same site enabling automatic download of changed BLAST databases. Thus it is possible to schedule a nightly update of changed databases and have the updated BLAST database stored locally or on a shared network drive at all times. Most BLAST databases on the NCBI site are updated on a daily basis to include all recent sequence submissions to GenBank.

A few commercial software packages are available for searching your own data. The advantage of using a commercial program is obvious when BLAST is integrated with the existing tools of these programs. Furthermore, they let you perform BLAST searches and retain annotations on the query sequence (see figure 12.22). It is also much easier to batch download a selection of hit sequences for further inspection.

Figure 12.22: *Snippet of alignment view of BLAST results from CLC Main Workbench. Individual alignments are represented directly in a graphical view. The top sequence is the query sequence and is shown with a selection of annotations.*

12.5.8 What you cannot get out of BLAST

Don't expect BLAST to produce the best available alignment. BLAST is a heuristic method which does not guarantee the best results, and therefore you cannot rely on BLAST if you wish to find *all* the hits in the database.

Instead, use the Smith-Waterman algorithm for obtaining the best possible local alignments [Smith and Waterman, 1981].

BLAST only makes local alignments. This means that a great but short hit in another sequence may not at all be related to the query sequence even though the sequences align well in a small region. It may be a domain or similar.

It is always a good idea to be cautious of the material in the database. For instance, the sequences may be wrongly annotated; hypothetical proteins are often simple translations of a found ORF on a sequenced nucleotide sequence and may not represent a true protein.

Don't expect to see the best result using the default settings. As described above, the settings should be adjusted according to the what kind of query sequence is used, and what kind of results you want. It is a good idea to perform the same BLAST search with different settings to get an idea of how they work. There is not a final answer on how to adjust the settings for your particular sequence.

12.5.9 Other useful resources

The BLAST web page hosted at NCBI
`http://www.ncbi.nlm.nih.gov/BLAST`

Download pages for the BLAST programs
`http://www.ncbi.nlm.nih.gov/BLAST/download.shtml`

Download pages for pre-formatted BLAST databases
`ftp://ftp.ncbi.nlm.nih.gov/blast/db/`

O'Reilly book on BLAST
`http://www.oreilly.com/catalog/blast/`

Explanation of scoring/substitution matrices and more
`http://www.clcbio.com/be/`

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

Chapter 13

3D molecule viewing

Contents

13.1 Importing structure files 186
13.2 Viewing structure files 187
13.2.1 Moving and rotating 187
13.3 Selections and display of the 3D structure 188
13.3.1 Coloring of the 3D structure 188
13.3.2 Hierarchical view - changing how selections of the structure are displayed 189
13.4 3D Output 193

In order to understand protein function it is often valuable to see the actual three dimensional structure of the protein. This is of course only possible if the structure of the protein has been resolved and published. *CLC Protein Workbench* has an integrated viewer of structure files. Structure files are usually deposited at the Protein DataBank (PDB) www.rcsb.org, where protein structure files can be searched and downloaded.

13.1 Importing structure files

In order to view the three dimensional structure files there are different ways to import these. The supported file formats are PDB and mmCIF which both can be downloaded from the Protein DataBank (http://www.rcsb.org) and imported through the import menu (see section 7.1.1).

Another way to import structure files is if a structure file is found through a direct search at the GenBank structure database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Structure). Read more about search for structures in section 11.3.

It is also possible to make a BLAST search against the PDB database. In the latter case, structure files can be directly downloaded to the navigation area by clicking the **Open structure** button below all the BLAST hits. Downloading structure files from a conducted BLAST search is only possible if the results are shown in a BLAST table. (See figure 13.1). How to conduct a BLAST search can be seen in section 12.1.1.

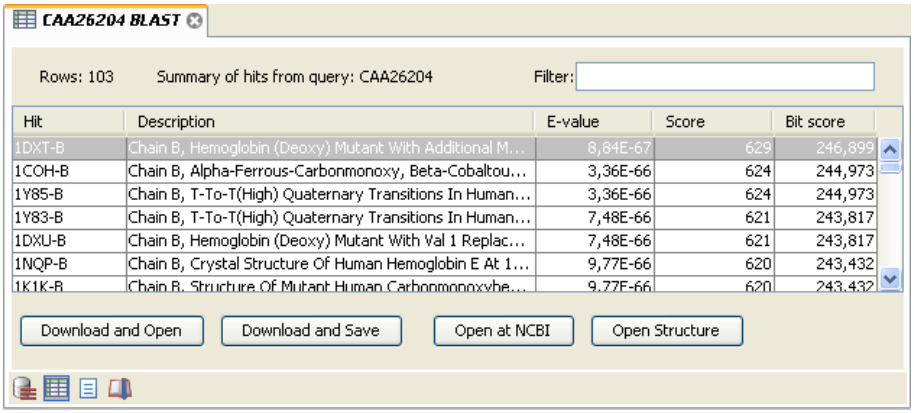


Figure 13.1: *It is possible to open a structure file directly from the output of a conducted BLAST search by clicking the Open Structure button.*

13.2 Viewing structure files

An example of a 3D structure is shown in figure 13.2.

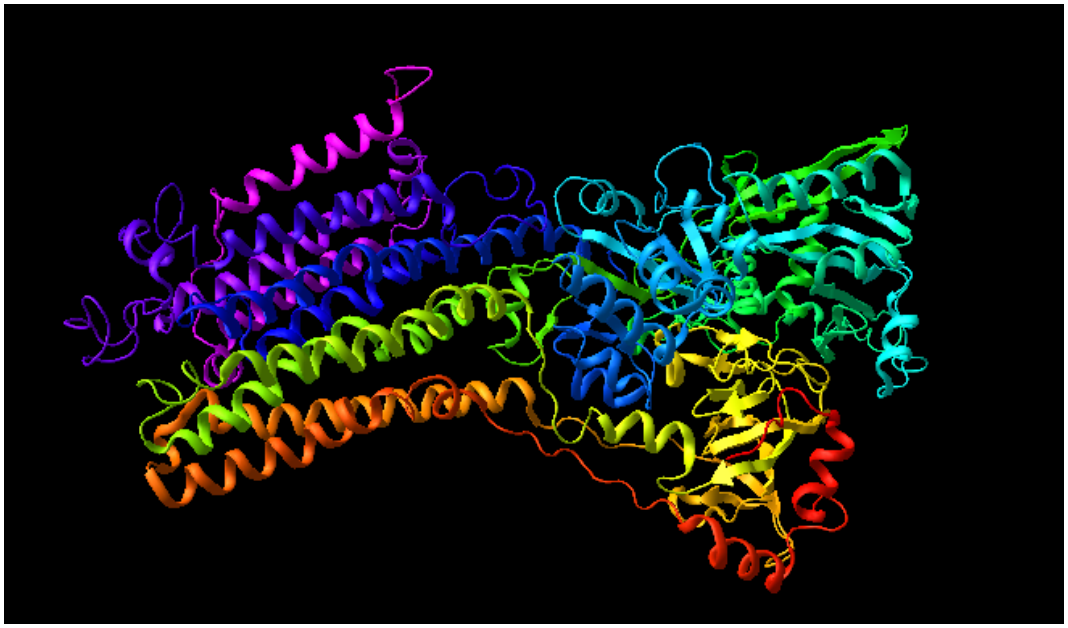


Figure 13.2: *3D view. Structure files can be opened, viewed and edited in several ways.*

Structures can be rotated and moved using the mouse and keyboard. **Pan mode** ( ) must be enabled in order to rotate and move the sequence.

Note! It is only possible to view one structure file at a time, in order to limit the amount of memory used.

13.2.1 Moving and rotating

Structure files are simply rotated by holding down the left mouse button while moving the mouse. This will rotate the structure in the direction the mouse is moved. The structures can be freely rotated in all directions.

Holding down the Ctrl on Windows or ⌘ on Mac key on the keyboard while dragging the mouse

moves the structure in the direction the mouse is moved. This is particularly useful if the view is zoomed to cover only a small region of the protein structure.

Zoom in ( ) and zoom out ( ) on the structure is done by selecting the appropriate zoom tool in the toolbar and clicking with the mouse on the view area. Alternatively, click and hold the left mouse button while using either zoom tool and move the mouse up or down to zoom out or in respectively. The view can be restored to display the entire structure by clicking the **Fit width** ( ) button on the toolbar (read more about zooming in section 3.3).

- **Rotate mode**
  The structure is rotated when the "Pan mode" ( ) is selected in the toolbar. If the "pan mode" is not enabled on the first view of a structure a warning is shown.
- **Zoom mode**
  Use the zoom buttons on the toolbar to enable zoom mode. A single click with the mouse will zoom slightly on the structure. Moreover, it is possible to zoom in and out on the structure by keeping the left mouse button pressed while moving the mouse up and down.
- **Move mode**
  It is possible to move the structure from side to side if the Ctrl key on Windows and ⌘ key on Mac is pressed while dragging with the mouse.

13.3 Selections and display of the 3D structure

The view of the structure can be changed in several ways. All graphical changes are carried out through the **Side Panel**. At the top, you can change the default coloring (**Default colors** and **Settings**), and at the bottom you can change the representation of specific parts of the structure in order to high-light e.g. an active site.

13.3.1 Coloring of the 3D structure

The default colors apply if nothing else has been specified in the **Hierarchical view** below (see section 13.3.2):

- **Atom type**. Colors the atoms individually.
  - **Carbon**: Light grey
  - **Oxygen**: Red
  - **Hydrogen**: White
  - **Nitrogen**: Light blue
  - **Sulphur**: Yellow
  - **Chlorine, Boron**: Green
  - **Phosphorus, Iron, Barium**: Orange
  - **Sodium**: Blue
  - **Magnesium**: Forest green
  - **Zn, Cu, Ni, Br**: Brown
  - **Ca, Mn, Al, Ti, Cr, Ag**: Dark grey

  - **F, Si, Au**: Goldenrod
  - **Iodine**: Purple
  - **Lithium**: firebrick
  - **Helium**: Pink
  - **Other**: Deep pink

- **Entities.** This will color protein subunits and additional structures individually. Using the view table, the user may select which colors are used to color subunits.
- **Rainbow.** This color mode will color the structure with rainbow colors along the sequence.
- **Residue hydrophobicity.** Colors the residues according to hydrophobicity.

In the **Settings** group, you can specify the background color to use. Default is black.

13.3.2 Hierarchical view - changing how selections of the structure are displayed

In the bottom of the **Side Panel**, you see the hierarchical view of the 3D structure (see an example in figure 13.3).

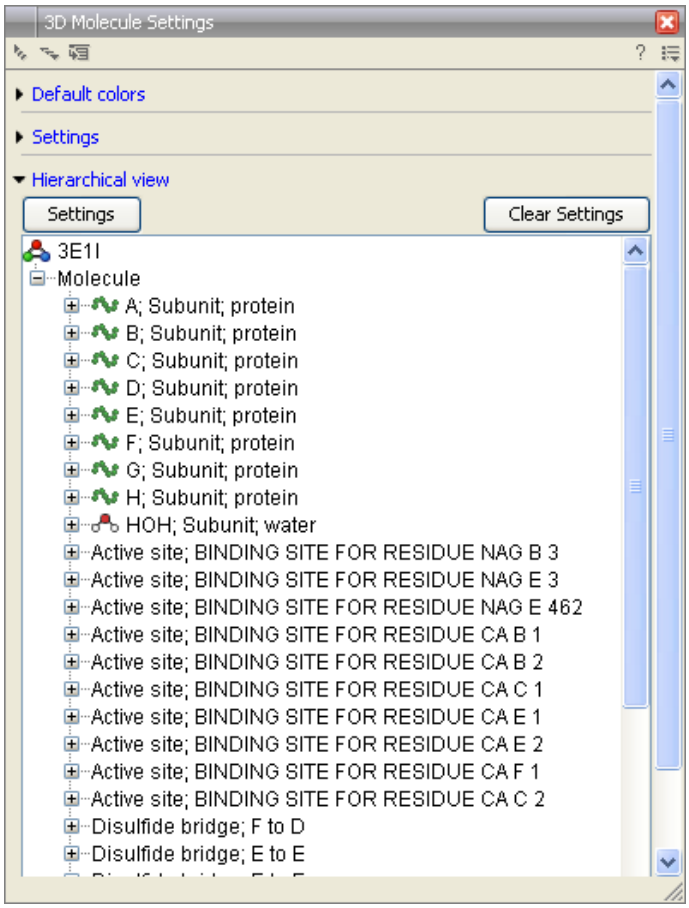

3D Molecule Settings
Default colors
Settings
Hierarchical view
Settings
Clear Settings
3E1I
Molecule
A; Subunit; protein
B; Subunit; protein
C; Subunit; protein
D; Subunit; protein
E; Subunit; protein
F; Subunit; protein
G; Subunit; protein
H; Subunit; protein
HOH; Subunit; water
Active site; BINDING SITE FOR RESIDUE NAG B 3
Active site; BINDING SITE FOR RESIDUE NAG E 3
Active site; BINDING SITE FOR RESIDUE NAG E 462
Active site; BINDING SITE FOR RESIDUE CA B 1
Active site; BINDING SITE FOR RESIDUE CA B 2
Active site; BINDING SITE FOR RESIDUE CA C 1
Active site; BINDING SITE FOR RESIDUE CA E 1
Active site; BINDING SITE FOR RESIDUE CA E 2
Active site; BINDING SITE FOR RESIDUE CA F 1
Active site; BINDING SITE FOR RESIDUE CA C 2
Disulfide bridge; F to D
Disulfide bridge; E to E

Figure 13.3: *Hierarchy view in the Side Panel.*

The hierarchical view shows the structure in a detailed manner. Individual structure subunits, residues, active sites, disulfide bridges or even down to the atom level can be selected individually and colored accordingly.

You can show additional information from the hierarchical view by holding your mouse still for one second (see an example in figure 13.4).

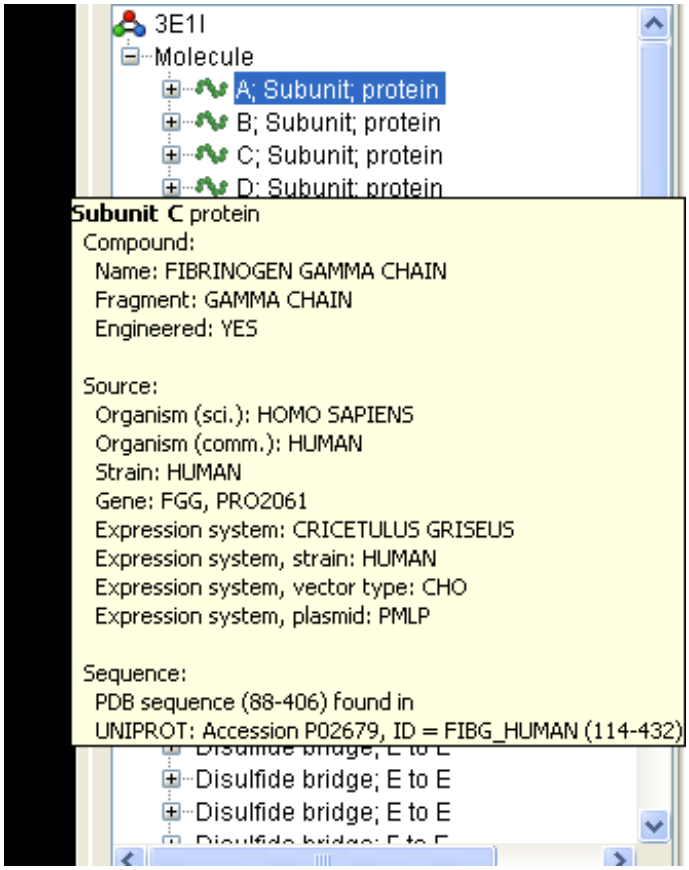


Figure 13.4: *Details shown by holding the mouse cursor on a subunit.*

For each item in the hierarchy view, you can apply special view settings. Simply click an item in the hierarchy view and click the **Settings** button above. This will display the settings dialog as shown in figure 13.5.

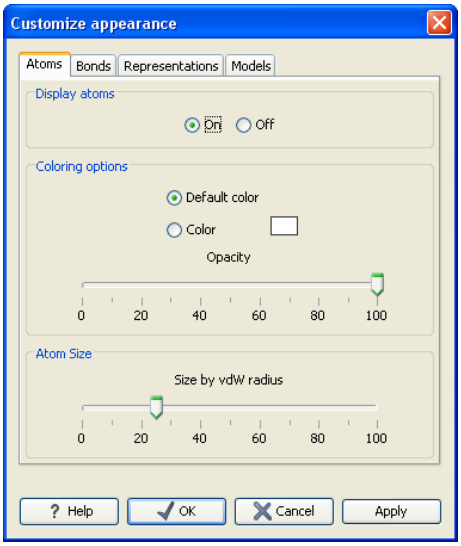


Figure 13.5: *Customize appearance.*

There are four tabs at the top. You can specify settings for each tab and then click **OK** to apply and close the dialog. You can also click **Apply** which will keep the settings dialog visible. You will then be able to select another item in the hierarchical view and apply settings for this also.

Atoms

The **Atoms** tab is shown in figure 13.5. At the top, you can choose to show atoms, and below you can specify their appearance:

- **Color**. Clicking the color box allows you to select a color.
- **Opacity**. Determines the level of opacity.
- **Atom size**. The size of the atoms measured by vdW radius.

Bonds

The **Bonds** tab is shown in figure 13.6.

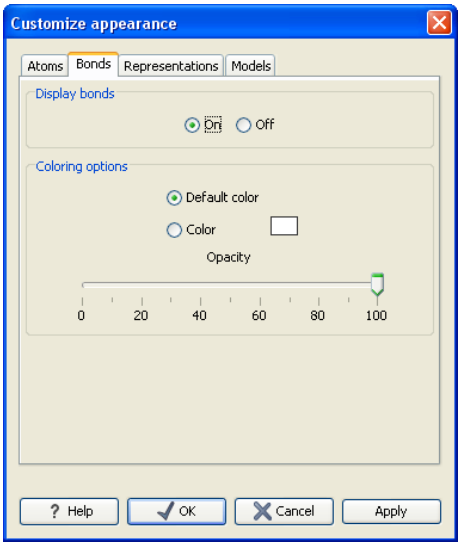


Figure 13.6: *Customize appearance for bonds.*

At the top, you can choose to show bonds, and below you can specify their appearance:

- **Color**. Clicking the color box allows you to select a color.
- **Opacity**. Determines the level of opacity.

Representations

The **Representations** tab is shown in figure 13.7.

At the top, you can choose to between four display models:

- **Secondary structure**.

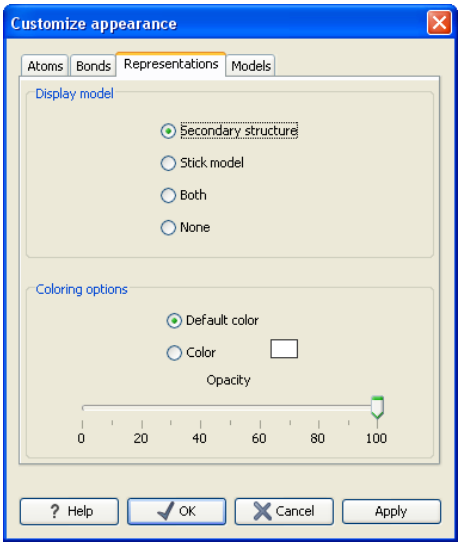


Figure 13.7: *Customize appearance for representations.*

- **Stick model**.
- **Both**. Displaying both secondary structure and stick model.
- **None**. Will not display representations.

- **Color**. Clicking the color box allows you to select a color.
- **Opacity**. Determines the level of opacity.

Models

The **Models** tab is shown in figure 13.8. At the top, you can choose to between three display

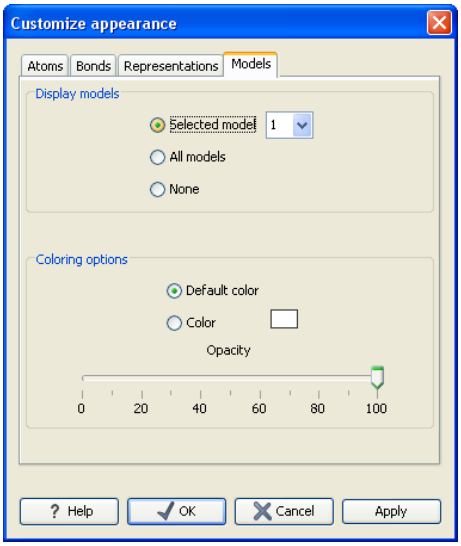


Figure 13.8: *Customize appearance for models.*

modes. This functionality is only applicable to NMR structures which have multiple resolved structures, for X-ray structures only one structure is available:

- **Selected model**.
- **All models**.
- **None**. Will not display models.

- **Color**. Clicking the color box allows you to select a color.
- **Opacity**. Determines the level of opacity.

13.4 3D Output

The output of the 3D view is rendered on the screen in real time and changes to the preferences are visible immediately. From *CLC Protein Workbench* you can export the visible part of the 3D view to different graphic formats, by pressing the **Graphics** button ( ) on the **Menu bar**. This will allow you to export in the following formats:

Format	Suffix	Type
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

Printing is not fully implemented with the 3D editor. Should you wish to print a 3D view, this can be done by either exporting to a graphics format and printing that or use the scheme below.

Windows:

- Adjust your 3D view in *CLC Protein Workbench*
- Press **Print Screen** on your keyboard (or **Alt + Print Screen**)
- Paste the result into an 'image editor' e.g. Paint or GIMP http://www.gimp.org/
- Crop (edit the screen shot)
- Save in your preferred file format and/or print

Mac:

- Set up your 3D view
- Press ⌘ **+ shift + 3** (or ⌘ **+ shift + 4**) (to take screen shot)
- Open the saved file (.pdf or .png) in a 'image editor' e.g. GIMP http://www.gimp.org/
- Crop (edit the screen shot)
- Save in your preferred file format and/or print

Linux:

- Set up your 3D view
- e.g. use GIMP to take the screen shot http://www.gimp.org/
- Crop (edit the screen shot)
- Save in your preferred file format and/or print

Chapter 14

General sequence analyses

Contents

14.1 Shuffle sequence . . . 195
14.2 Dot plots . . . 197
14.2.1 Create dot plots . . . 197
14.2.2 View dot plots . . . 199
14.2.3 Bioinformatics explained: Dot plots . . . 200
14.2.4 Bioinformatics explained: Scoring matrices . . . 204
14.3 Local complexity plot . . . 207
14.4 Sequence statistics . . . 208
14.4.1 Bioinformatics explained: Protein statistics . . . 211
14.5 Join sequences . . . 214
14.6 Pattern Discovery . . . 215
14.6.1 Pattern discovery search parameters . . . 216
14.6.2 Pattern search output . . . 217
14.7 Motif Search . . . 217
14.7.1 Dynamic motifs . . . 218
14.7.2 Motif search from the Toolbox . . . 220
14.7.3 Java regular expressions . . . 221
14.7.4 Create motif list . . . 223

CLC Protein Workbench offers different kinds of sequence analyses, which apply to both protein and DNA. The analyses are described in this chapter.

14.1 Shuffle sequence

In some cases, it is beneficial to shuffle a sequence. This is an option in the **Toolbox** menu under **General Sequence Analyses**. It is normally used for statistical analyses, e.g. when comparing an alignment score with the distribution of scores of shuffled sequences.

Shuffling a sequence removes all annotations that relate to the residues.

select sequence | Toolbox in the Menu Bar | General Sequence Analyses ( ) | Shuffle Sequence ( )

or **right-click a sequence** | **Toolbox** | **General Sequence Analyses** ( ) | **Shuffle Sequence** ( )

This opens the dialog displayed in figure 14.1:

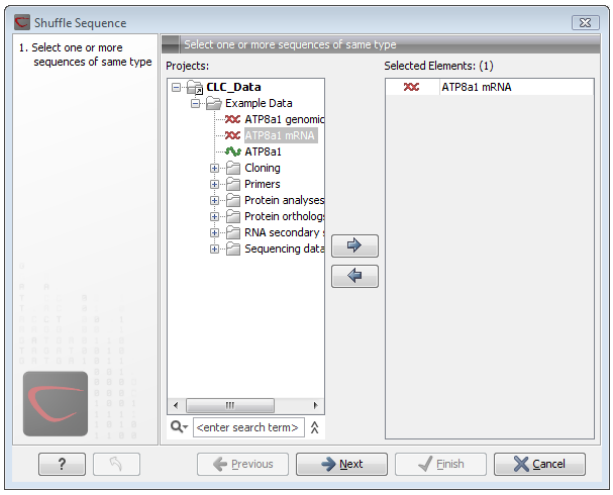


Figure 14.1: *Choosing sequence for shuffling.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

Click **Next** to determine how the shuffling should be performed.

In this step, shown in figure 14.2: For nucleotides, the following parameters can be set:

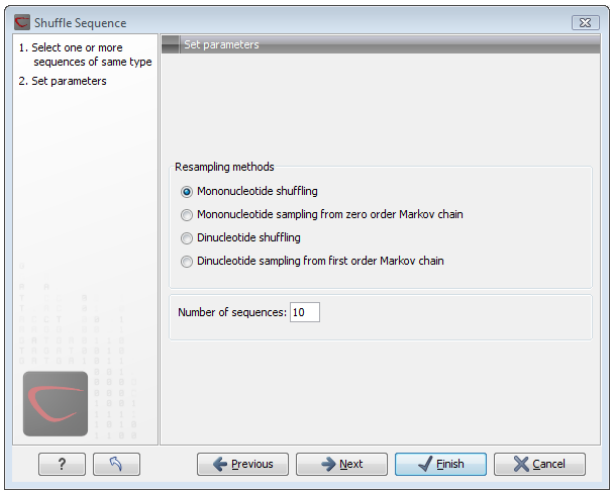


Figure 14.2: *Parameters for shuffling.*

- **Mononucleotide shuffling.** Shuffle method generating a sequence of the exact same mononucleotide frequency
- **Dinucleotide shuffling.** Shuffle method generating a sequence of the exact same dinucleotide frequency
- **Mononucleotide sampling from zero order Markov chain.** Resampling method generating a sequence of the same expected mononucleotide frequency.

- **Dinucleotide sampling from first order Markov chain.** Resampling method generating a sequence of the same expected dinucleotide frequency.

For proteins, the following parameters can be set:

- **Single amino acid shuffling.** Shuffle method generating a sequence of the exact same amino acid frequency.
- **Single amino acid sampling from zero order Markov chain.** Resampling method generating a sequence of the same expected single amino acid frequency.
- **Dipeptide shuffling.** Shuffle method generating a sequence of the exact same dipeptide frequency.
- **Dipeptide sampling from first order Markov chain.** Resampling method generating a sequence of the same expected dipeptide frequency.

For further details of these algorithms, see [Clote et al., 2005]. In addition to the shuffle method, you can specify the number of randomized sequences to output.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the shuffled sequence. The new sequence is not saved automatically. To save the sequence, drag it into the **Navigation Area** or press ctrl + S (⌘ + S on Mac) to activate a save dialog.

14.2 Dot plots

Dot plots provide a powerful visual comparison of two sequences. Dot plots can also be used to compare regions of similarity within a sequence. This chapter first describes how to create and second how to adjust the view of the plot.

14.2.1 Create dot plots

A dot plot is a simple, yet intuitive way of comparing two sequences, either DNA or protein, and is probably the oldest way of comparing two sequences [Maizel and Lenk, 1981]. A dot plot is a 2 dimensional matrix where each axis of the plot represents one sequence. By sliding a fixed size window over the sequences and making a sequence match by a dot in the matrix, a diagonal line will emerge if two identical (or very homologous) sequences are plotted against each other. Dot plots can also be used to visually inspect sequences for direct or inverted repeats or regions with low sequence complexity. Various smoothing algorithms can be applied to the dot plot calculation to avoid noisy background of the plot. Moreover, can various substitution matrices be applied in order to take the evolutionary distance of the two sequences into account.

To create a dot plot:

Toolbox | **General Sequence Analyses** ( ) | **Create Dot Plot** ( )

or **Select one or two sequences in the Navigation Area** | **Toolbox in the Menu Bar** | **General Sequence Analyses** ( ) | **Create Dot Plot** ( )

or **Select one or two sequences in the Navigation Area** | **right-click in the Navigation Area** | **Toolbox** | **General Sequence Analyses** ( ) | **Create Dot Plot** ( )

This opens the dialog shown in figure 14.3.

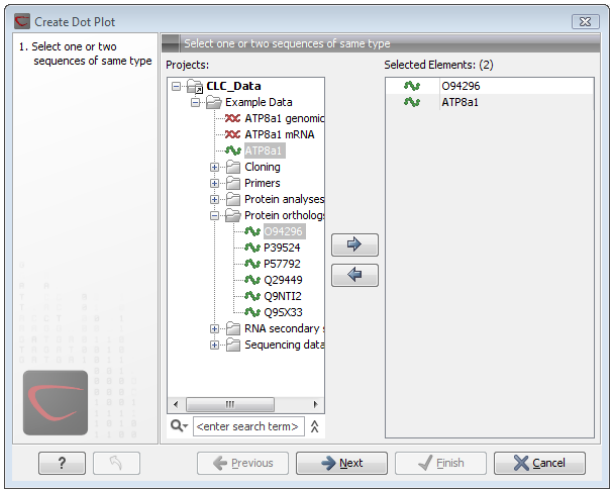


Figure 14.3: *Selecting sequences for the dot plot.*

If a sequence was selected before choosing the **Toolbox** action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove elements from the selected elements. Click **Next** to adjust dot plot parameters. Clicking **Next** opens the dialog shown in figure 14.4.

Notice! Calculating dot plots take up a considerable amount of memory in the computer. Therefore, you see a warning if the sum of the number of nucleotides/amino acids in the sequences is higher than 8000. If you insist on calculating a dot plot with more residues the Workbench may shut down, allowing you to save your work first. However, this depends on your computer's memory configuration.

Adjust dot plot parameters

There are two parameters for calculating the dot plot:

- **Distance correction (only valid for protein sequences)** In order to treat evolutionary transitions of amino acids, a distance correction measure can be used when calculating the dot plot. These distance correction matrices (substitution matrices) take into account the likeliness of one amino acid changing to another.
- **Window size** A residue by residue comparison (window size = 1) would undoubtedly result in a very noisy background due to a lot of similarities between the two sequences of interest. For DNA sequences the background noise will be even more dominant as a match between only four nucleotide is very likely to happen. Moreover, a residue by residue comparison (window size = 1) can be very time consuming and computationally demanding. Increasing the window size will make the dot plot more 'smooth'.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

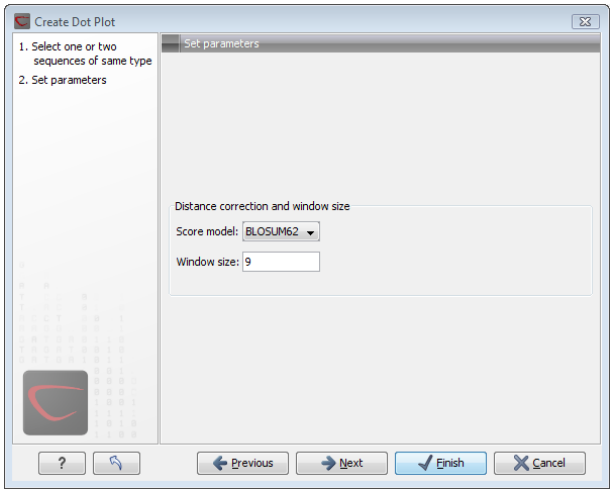


Figure 14.4: *Setting the dot plot parameters.*

14.2.2 View dot plots

A view of a dot plot can be seen in figure 14.5. You can select **Zoom in** ( ) in the Toolbar and click the dot plot to zoom in to see the details of particular areas.

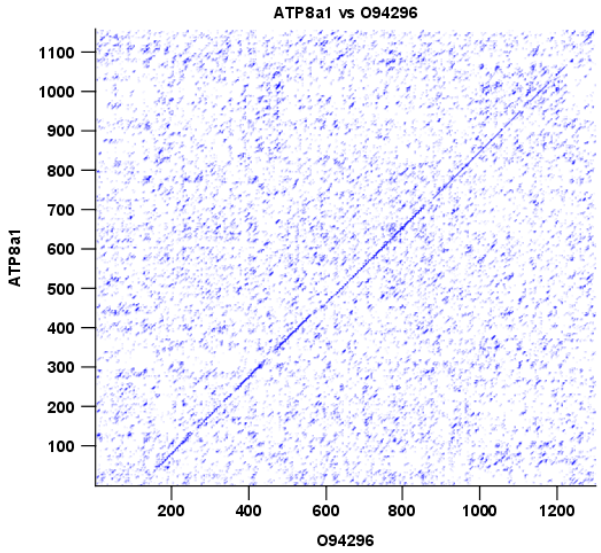


Figure 14.5: *A view is opened showing the dot plot.*

The **Side Panel** to the right let you specify the dot plot preferences. The gradient color box can be adjusted to get the appropriate result by dragging the small pointers at the top of the box. Moving the slider from the right to the left lowers the thresholds which can be directly seen in the dot plot, where more diagonal lines will emerge. You can also choose another color gradient by clicking on the gradient box and choose from the list.

Adjusting the sliders above the gradient box is also practical, when producing an output for printing. (Too much background color might not be desirable). By crossing one slider over the other (the two sliders change side) the colors are inverted, allowing for a white background. (If you choose a color gradient, which includes white). Se figure 14.5.

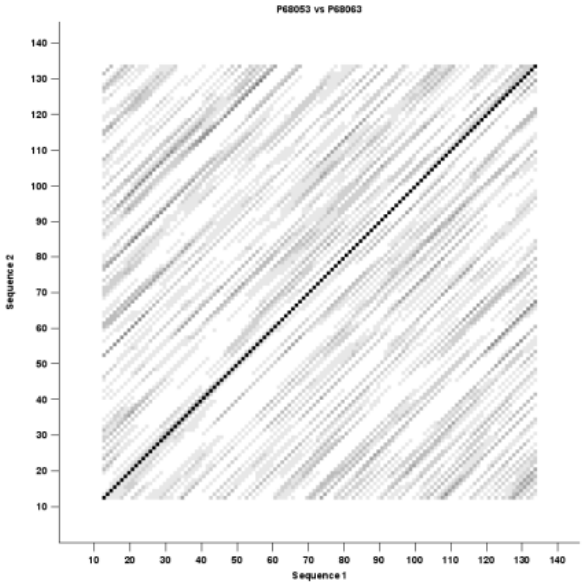


Figure 14.6: *Dot plot with inverted colors, practical for printing.*

14.2.3 Bioinformatics explained: Dot plots

Realization of dot plots

Dot plots are two-dimensional plots where the x-axis and y-axis each represents a sequence and the plot itself shows a comparison of these two sequences by a calculated score for each position of the sequence. If a window of fixed size on one sequence (one axis) match to the other sequence a dot is drawn at the plot. Dot plots are one of the oldest methods for comparing two sequences [Maizel and Lenk, 1981].

The scores that are drawn on the plot are affected by several issues.

- Scoring matrix for distance correction.
  Scoring matrices (BLOSUM and PAM) contain substitution scores for every combination of two amino acids. Thus, these matrices can only be used for dot plots of protein sequences.
- Window size
  The single residue comparison (bit by bit comparison(window size = 1)) in dot plots will undoubtedly result in a noisy background of the plot. You can imagine that there are many successes in the comparison if you only have four possible residues like in nucleotide sequences. Therefore you can set a window size which is smoothing the dot plot. Instead of comparing single residues it compares subsequences of length set as window size. The score is now calculated with respect to aligning the subsequences.
- Threshold
  The dot plot shows the calculated scores with colored threshold. Hence you can better recognize the most important similarities.

Examples and interpretations of dot plots

Contrary to simple sequence alignments dot plots can be a very useful tool for spotting various evolutionary events which may have happened to the sequences of interest.

Below is shown some examples of dot plots where sequence insertions, low complexity regions, inverted repeats etc. can be identified visually.

Similar sequences

The most simple example of a dot plot is obtained by plotting two homologous sequences of interest. If very similar or identical sequences are plotted against each other a diagonal line will occur.

The dot plot in figure 14.7 shows two related sequences of the Influenza A virus nucleoproteins infecting ducks and chickens. Accession numbers from the two sequences are: DQ232610 and DQ023146. Both sequences can be retrieved directly from `http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi`.

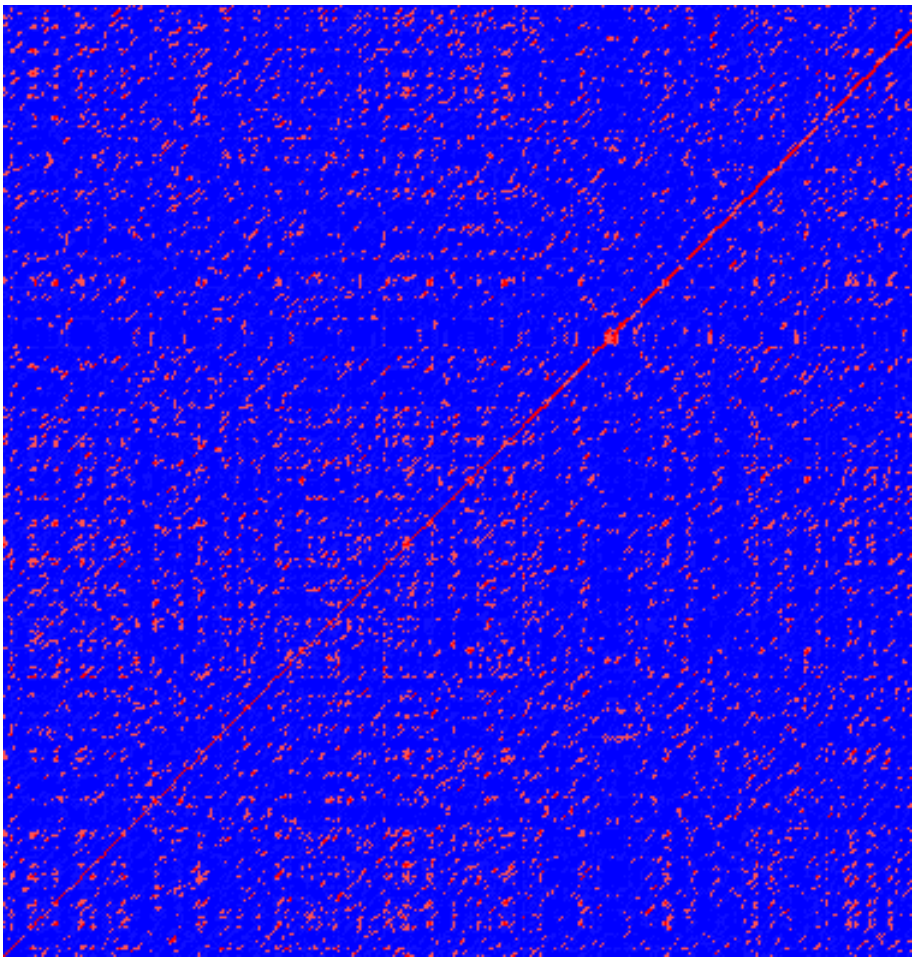


Figure 14.7: *Dot plot of DQ232610 vs. DQ023146 (Influenza A virus nucleoproteins) showing and overall similarity*

Repeated regions

Sequence repeats can also be identified using dot plots. A repeat region will typically show up as lines parallel to the diagonal line.

If the dot plot shows more than one diagonal in the same region of a sequence, the regions depending to the other sequence are repeated. In figure 14.9 you can see a sequence with repeats.

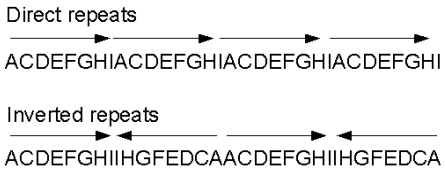


Figure 14.8: *Direct and inverted repeats shown on an amino acid sequence generated for demonstration purposes.*

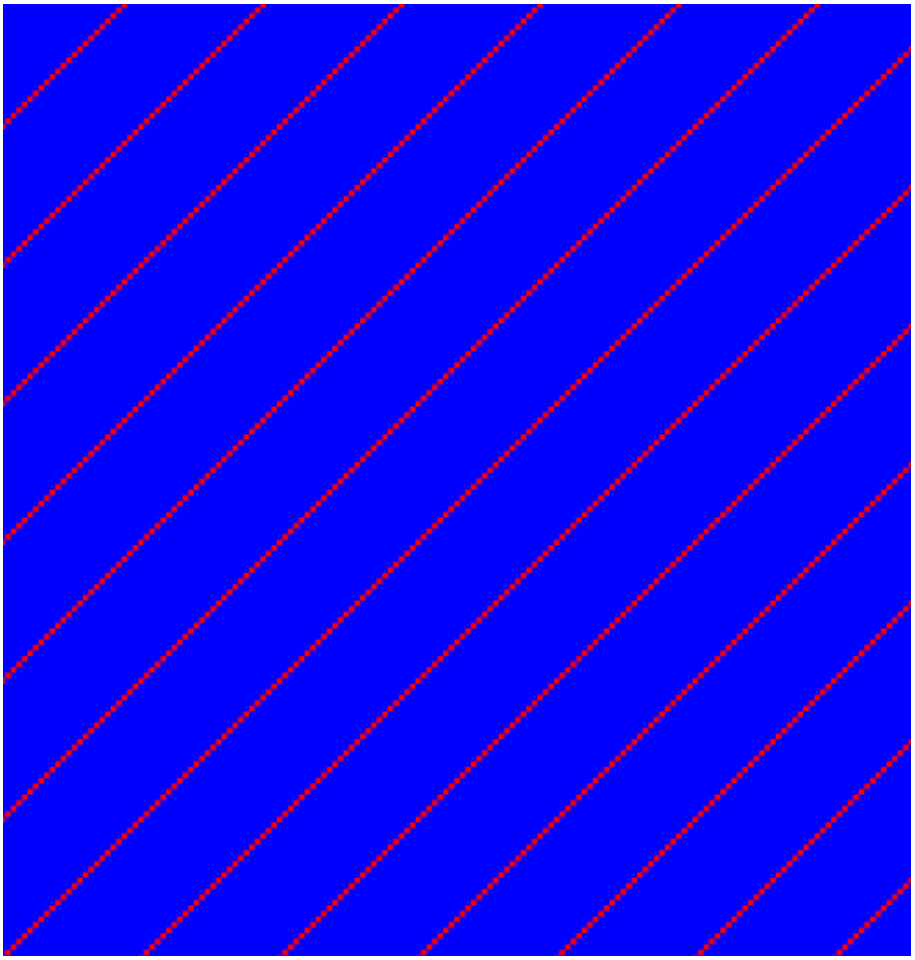


Figure 14.9: *The dot plot of a sequence showing repeated elements. See also figure 14.8.*

Frame shifts

Frame shifts in a nucleotide sequence can occur due to insertions, deletions or mutations. Such frame shifts can be visualized in a dot plot as seen in figure 14.10. In this figure, three frame shifts for the sequence on the y-axis are found.

1. Deletion of nucleotides
2. Insertion of nucleotides
3. Mutation (out of frame)

Sequence inversions

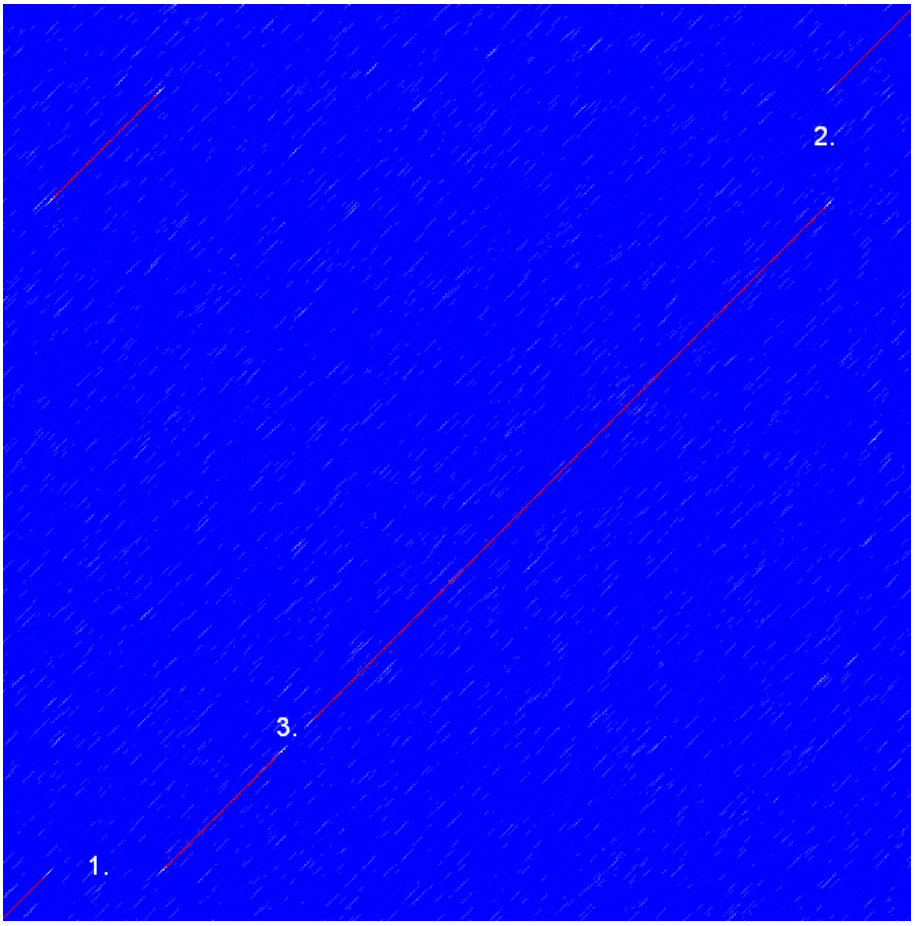


Figure 14.10: *This dot plot show various frame shifts in the sequence. See text for details.*

In dot plots you can see an inversion of sequence as contrary diagonal to the diagonal showing similarity. In figure 14.11 you can see a dot plot (window length is 3) with an inversion.

Low-complexity regions

Low-complexity regions in sequences can be found as regions around the diagonal all obtaining a high score. Low complexity regions are calculated from the redundancy of amino acids within a limited region [Wootton and Federhen, 1993]. These are most often seen as short regions of only a few different amino acids. In the middle of figure 14.12 is a square shows the low-complexity region of this sequence.

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

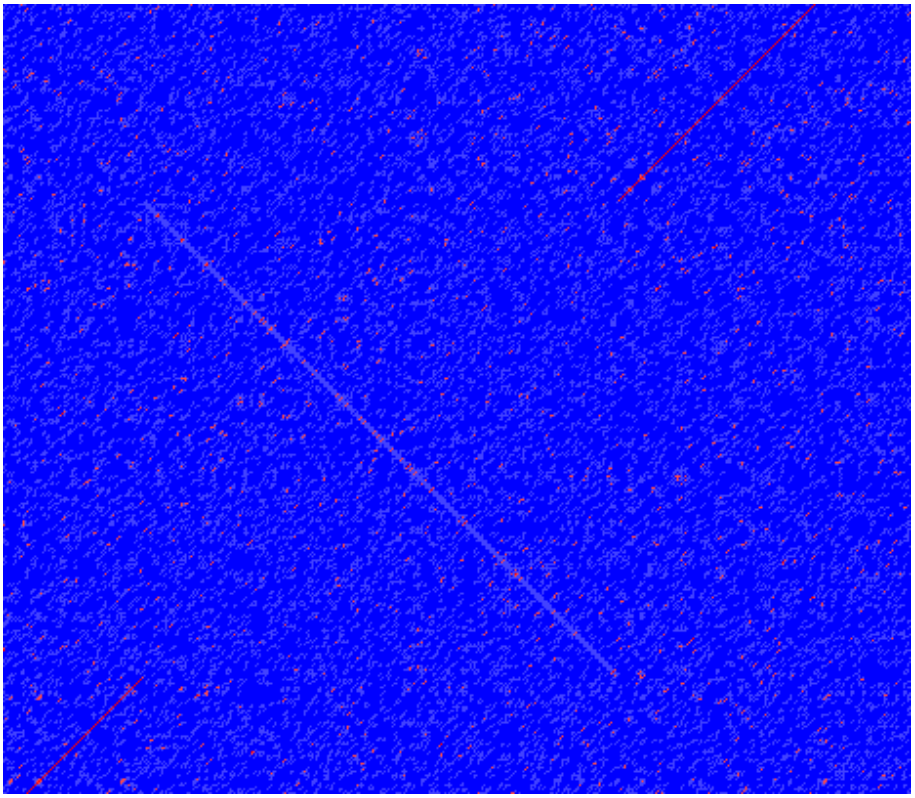


Figure 14.11: *The dot plot showing a inversion in a sequence. See also figure 14.8.*

SOME RIGHTS RESERVED

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

14.2.4 Bioinformatics explained: Scoring matrices

Biological sequences have evolved throughout time and evolution has shown that not all changes to a biological sequence is equally likely to happen. Certain amino acid substitutions (change of one amino acid to another) happen often, whereas other substitutions are very rare. For instance, tryptophan (W) which is a relatively rare amino acid, will only — on very rare occasions — mutate into a leucine (L).

Based on evolution of proteins it became apparent that these changes or substitutions of amino acids can be modeled by a scoring matrix also refereed to as a substitution matrix. See an example of a scoring matrix in table 14.1. This matrix lists the substitution scores of every single amino acid. A score for an aligned amino acid pair is found at the intersection of the corresponding column and row. For example, the substitution score from an arginine (R) to a lysine (K) is 2. The diagonal show scores for amino acids which have not changed. Most substitutions changes have a negative score. Only rounded numbers are found in this matrix.

The two most used matrices are the BLOSUM [Henikoff and Henikoff, 1992] and PAM [Dayhoff and Schwartz, 1978].

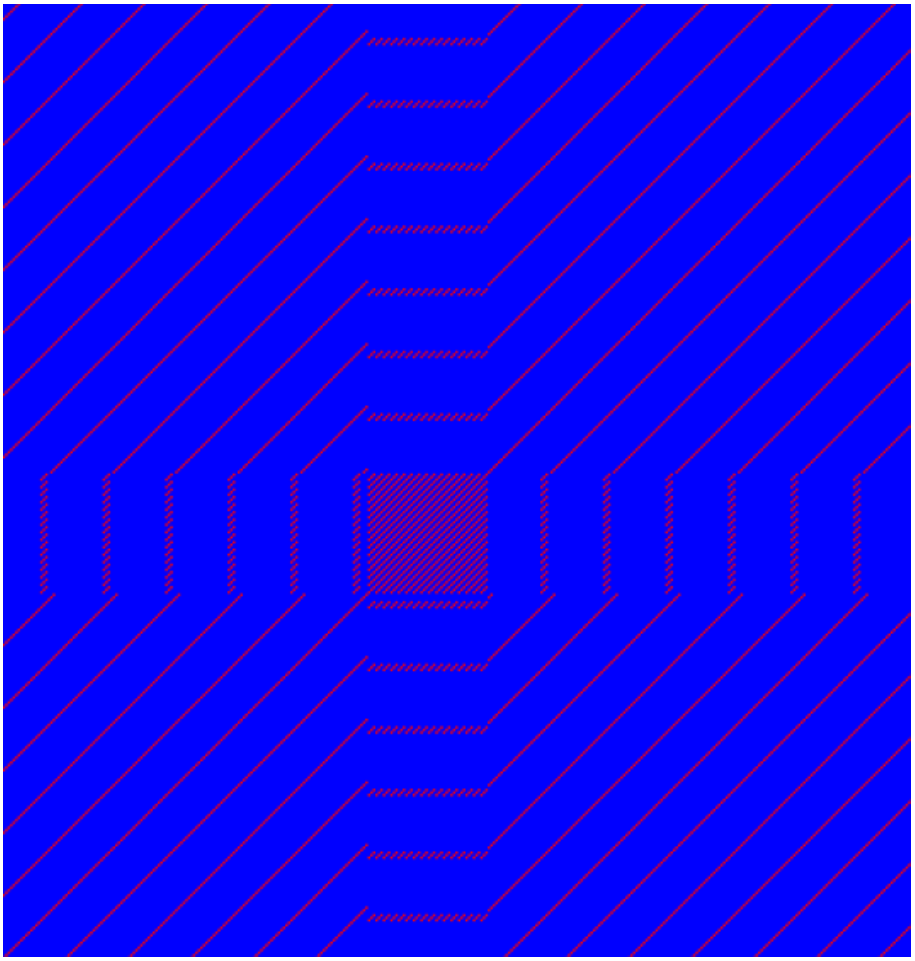


Figure 14.12: *The dot plot showing a low-complexity region in the sequence. The sequence is artificial and low complexity regions does not always show as a square.*

Different scoring matrices

PAM

The first PAM matrix (Point Accepted Mutation) was published in 1978 by Dayhoff et al. The PAM matrix was build through a global alignment of related sequences all having sequence similarity above 85% [Dayhoff and Schwartz, 1978]. A PAM matrix shows the probability that any given amino acid will mutate into another in a given time interval. As an example, PAM1 gives that one amino acid out of a 100 will mutate in a given time interval. In the other end of the scale, a PAM256 matrix, gives the probability of 256 mutations in a 100 amino acids (see figure 14.13).

There are some limitation to the PAM matrices which makes the BLOSUM matrices somewhat more attractive. The dataset on which the initial PAM matrices were build is very old by now, and the PAM matrices assume that all amino acids mutate at the same rate - this is not a correct assumption.

BLOSUM

In 1992, 14 years after the PAM matrices were published, the BLOSUM matrices (BLOcks SUbstitution Matrix) were developed and published [Henikoff and Henikoff, 1992].

Henikoff et al. wanted to model more divergent proteins, thus they used locally aligned sequences where none of the aligned sequences share less than 62% identity. This resulted

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Table 14.1: **The BLOSUM62 matrix**. A tabular view of the BLOSUM62 matrix containing all possible substitution scores [Henikoff and Henikoff, 1992].

in a scoring matrix´called BLOSUM62. In contrast to the PAM matrices the BLOSUM matrices are calculated from alignments without gaps emerging from the BLOCKS database `http://blocks.fhcrc.org/`.

Sean Eddy recently wrote a paper reviewing the BLOSUM62 substitution matrix and how to calculate the scores [Eddy, 2004].

Use of scoring matrices

Deciding which scoring matrix you should use in order of obtain the best alignment results is a difficult task. If you have no prior knowledge on the sequence the BLOSUM62 is probably the best choice. This matrix has become the *de facto* standard for scoring matrices and is also used as the default matrix in BLAST searches. The selection of a "wrong" scoring matrix will most probable strongly influence on the outcome of the analysis. In general a few rules apply to the selection of scoring matrices.

- For closely related sequences choose BLOSUM matrices created for highly similar alignments, like BLOSUM80. You can also select low PAM matrices such as PAM1.
- For distant related sequences, select low BLOSUM matrices (for example BLOSUM45) or high PAM matrices such as PAM250.

The BLOSUM matrices with low numbers correspond to PAM matrices with high numbers. (See figure 14.13) for correlations between the PAM and BLOSUM matrices. To summarize, if you want to find distant related proteins to a sequence of interest using BLAST, you could benefit of using BLOSUM45 or similar matrices.

Other useful resources

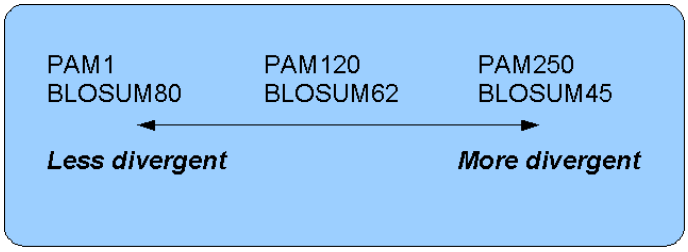


Figure 14.13: *Relationship between scoring matrices. The BLOSUM62 has become a* de facto *standard scoring matrix for a wide range of alignment programs. It is the default matrix in BLAST.*

Calculate your own PAM matrix
http://www.bioinformatics.nl/tools/pam.html

BLOKS database
http://blocks.fhcrc.org/

NCBI help site
http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/Scoring2.html

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

14.3 Local complexity plot

In *CLC Protein Workbench* it is possible to calculate local complexity for both DNA and protein sequences. The local complexity is a measure of the diversity in the composition of amino acids within a given range (window) of the sequence. The K2 algorithm is used for calculating local complexity [Wootton and Federhen, 1993]. To conduct a complexity calculation do the following:

> **Select sequences in Navigation Area** | **Toolbox in Menu Bar** | **General Sequence Analyses** ( ) | **Create Complexity Plot** ( )

This opens a dialog. In **Step 1** you can change, remove and add DNA and protein sequences.

When the relevant sequences are selected, clicking **Next** takes you to **Step 2**. This step allows you to adjust the window size from which the complexity plot is calculated. Default is set to 11 amino acids and the number should always be odd. The higher the number, the less volatile the graph.

Figure 14.14 shows an example of a local complexity plot.

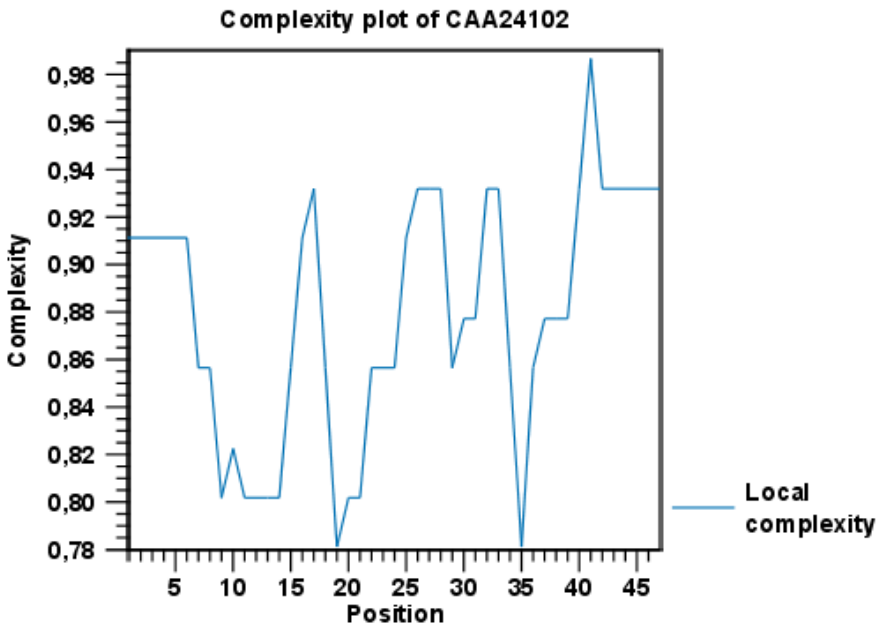


Figure 14.14: *An example of a local complexity plot.*

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. The values of the complexity plot approaches 1.0 as the distribution of amino acids become more complex.

See section B in the appendix for information about the graph view.

14.4 Sequence statistics

CLC Protein Workbench can produce an output with many relevant statistics for protein sequences. Some of the statistics are also relevant to produce for DNA sequences. Therefore, this section deals with both types of statistics. The required steps for producing the statistics are the same.

To create a statistic for the sequence, do the following:

> **select sequence(s)** | **Toolbox in the Menu Bar** | **General Sequence Analyses ( )** | **Create Sequence Statistics ( )**

This opens a dialog where you can alter your choice of sequences which you want to create statistics for. You can also add sequence lists.

Note! You cannot create statistics for DNA and protein sequences at the same time.

When the sequences are selected, click **Next**.

This opens the dialog displayed in figure 14.15.

The dialog offers to adjust the following parameters:

- **Individual statistics layout.** If more sequences were selected in **Step 1**, this function generates separate statistics for each sequence.
- **Comparative statistics layout.** If more sequences were selected in **Step 1**, this function generates statistics with comparisons between the sequences.

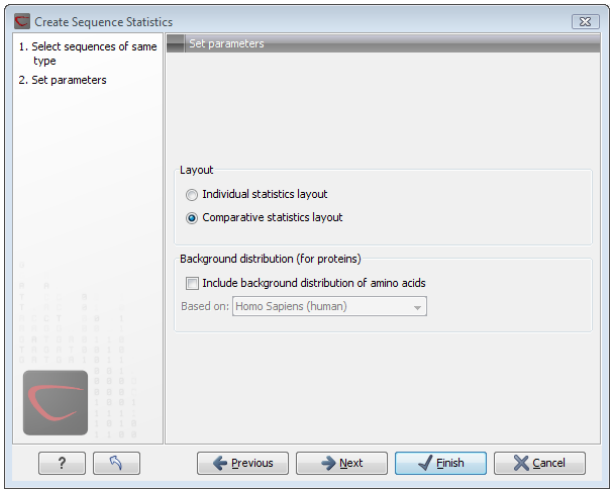


Figure 14.15: *Setting parameters for the sequence statistics.*

You can also choose to include Background distribution of amino acids. If this box is ticked, an extra column with amino acid distribution of the chosen species, is included in the table output. (The distributions are calculated from UniProt `www.uniprot.org` version 6.0, dated September 13 2005.)

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. An example of protein sequence statistics is shown in figure 14.16.

1 Protein statistics

1.1 Sequence information

Sequence type	Protein
Length	147
Organism	Mus musculus
Name	CAA32220
Description	haemoglobin beta-h0 chain [Mus musculus].
Modification Date	18-APR-2005
Weight	16,412 kDa

1.2 Half-life

N-terminal aa	Half-life mammals	Half-life yeast	Half-life E.Coli

Figure 14.16: *Comparative sequence statistics.*

Nucleotide sequence statistics are generated using the same dialog as used for protein sequence statistics. However, the output of Nucleotide sequence statistics is less extensive than that of the protein sequence statistics.

Note! The headings of the tables change depending on whether you calculate 'individual' or 'comparative' sequence statistics.

The output of comparative protein sequence statistics include:

- Sequence information:
  - Sequence type
  - Length
  - Organism

  - Name
  - Description
  - Modification Date
  - Weight. This is calculated like this: $sum_{unitsinsequence}(weight(unit)) - links * weight(H2O)$ where `links` is the sequence length minus one and `units` are amino acids. The atomic composition is defined the same way.
  - Isoelectric point
  - Aliphatic index
- Half-life
- Extinction coefficient
- Counts of Atoms
- Frequency of Atoms
- Count of hydrophobic and hydrophilic residues
- Frequencies of hydrophobic and hydrophilic residues
- Count of charged residues
- Frequencies of charged residues
- Amino acid distribution
- Histogram of amino acid distribution
- Annotation table
- Counts of di-peptides
- Frequency of di-peptides

The output of nucleotide sequence statistics include:

- General statistics:
  - Sequence type
  - Length
  - Organism
  - Name
  - Description
  - Modification Date
  - Weight. This is calculated like this: $sum_{unitsinsequence}(weight(unit)) - links * weight(H2O)$ where `links` is the sequence length minus one for linear sequences and sequence length for circular molecules. The `units` are monophosphates. Both the weight for single- and double stranded molecules are includes. The atomic composition is defined the same way.

- Atomic composition
- Nucleotide distribution table
- Nucleotide distribution histogram
- Annotation table
- Counts of di-nucleotides
- Frequency of di-nucleotides

A short description of the different areas of the statistical output is given in section 14.4.1.

14.4.1 Bioinformatics explained: Protein statistics

Every protein holds specific and individual features which are unique to that particular protein. Features such as isoelectric point or amino acid composition can reveal important information of a novel protein. Many of the features described below are calculated in a simple way.

Molecular weight

The molecular weight is the mass of a protein or molecule. The molecular weight is simply calculated as the sum of the atomic mass of all the atoms in the molecule.

The weight of a protein is usually represented in Daltons (Da).

A calculation of the molecular weight of a protein does not usually include additional posttranslational modifications. For native and unknown proteins it tends to be difficult to assess whether posttranslational modifications such as glycosylations are present on the protein, making a calculation based solely on the amino acid sequence inaccurate. The molecular weight can be determined very accurately by mass-spectrometry in a laboratory.

Isoelectric point

The isoelectric point (pI) of a protein is the pH where the proteins has no net charge. The pI is calculated from the pKa values for 20 different amino acids. At a pH below the pI, the protein carries a positive charge, whereas if the pH is above pI the proteins carry a negative charge. In other words, pI is high for basic proteins and low for acidic proteins. This information can be used in the laboratory when running electrophoretic gels. Here the proteins can be separated, based on their isoelectric point.

Aliphatic index

The aliphatic index of a protein is a measure of the relative volume occupied by aliphatic side chain of the following amino acids: alanine, valine, leucine and isoleucine. An increase in the aliphatic index increases the thermostability of globular proteins. The index is calculated by the following formula.

$$Aliphaticindex = X(Ala) + a * X(Val) + b * X(Leu) + b * (X)Ile$$

Amino acid	Mammalian	Yeast	E. coli
Ala (A)	4.4 hour	>20 hours	>10 hours
Cys (C)	1.2 hours	>20 hours	>10 hours
Asp (D)	1.1 hours	3 min	>10 hours
Glu (E)	1 hour	30 min	>10 hours
Phe (F)	1.1 hours	3 min	2 min
Gly (G)	30 hours	>20 hours	>10 hours
His (H)	3.5 hours	10 min	>10 hours
Ile (I)	20 hours	30 min	>10 hours
Lys (K)	1.3 hours	3 min	2 min
Leu (L)	5.5 hours	3 min	2 min
Met (M)	30 hours	>20 hours	>10 hours
Asn (N)	1.4 hours	3 min	>10 hours
Pro (P)	>20 hours	>20 hours	?
Gln (Q)	0.8 hour	10 min	>10 hours
Arg (R)	1 hour	2 min	2 min
Ser (S)	1.9 hours	>20 hours	>10 hours
Thr (T)	7.2 hours	>20 hours	>10 hours
Val (V)	100 hours	>20 hours	>10 hours
Trp (W)	2.8 hours	3 min	2 min
Tyr (Y)	2.8 hours	10 min	2 min

Table 14.2: **Estimated half life**. Half life of proteins where the N-terminal residue is listed in the first column and the half-life in the subsequent columns for mammals, yeast and *E. coli*.

X(Ala), *X(Val)*, *X(Ile)* and *X(Leu)* are the amino acid compositional fractions. The constants a and b are the relative volume of valine (a=2.9) and leucine/isoleucine (b=3.9) side chains compared to the side chain of alanine [Ikai, 1980].

Estimated half-life

The half life of a protein is the time it takes for the protein pool of that particular protein to be reduced to the half. The half life of proteins is highly dependent on the presence of the N-terminal amino acid, thus overall protein stability [Bachmair et al., 1986, Gonda et al., 1989, Tobias et al., 1991]. The importance of the N-terminal residues is generally known as the 'N-end rule'. The N-end rule and consequently the N-terminal amino acid, simply determines the half-life of proteins. The estimated half-life of proteins have been investigated in mammals, yeast and *E. coli* (see Table 14.2). If leucine is found N-terminally in mammalian proteins the estimated half-life is 5.5 hours.

Extinction coefficient

This measure indicates how much light is absorbed by a protein at a particular wavelength. The extinction coefficient is measured by UV spectrophotometry, but can also be calculated. The amino acid composition is important when calculating the extinction coefficient. The extinction coefficient is calculated from the absorbance of cysteine, tyrosine and tryptophan using the following equation:

$$Ext(Protein) = count(Cystine) * Ext(Cystine) + count(Tyr) * Ext(Tyr) + count(Trp) * Ext(Trp)$$

where Ext is the extinction coefficient of amino acid in question. At 280nm the extinction coefficients are: Cys=120, Tyr=1280 and Trp=5690.

This equation is only valid under the following conditions:

- pH 6.5
- 6.0 M guanidium hydrochloride
- 0.02 M phosphate buffer

The extinction coefficient values of the three important amino acids at different wavelengths are found in [Gill and von Hippel, 1989].

Knowing the extinction coefficient, the absorbance (optical density) can be calculated using the following formula:

$$Absorbance(Protein) = \frac{Ext(Protein)}{Molecular\ weight}$$

Two values are reported. The first value is computed assuming that all cysteine residues appear as half cystines, meaning they form di-sulfide bridges to other cysteines. The second number assumes that no di-sulfide bonds are formed.

Atomic composition

Amino acids are indeed very simple compounds. All 20 amino acids consist of combinations of only five different atoms. The atoms which can be found in these simple structures are: Carbon, Nitrogen, Hydrogen, Sulfur, Oxygen. The atomic composition of a protein can for example be used to calculate the precise molecular weight of the entire protein.

Total number of negatively charged residues (Asp+Glu)

At neutral pH, the fraction of negatively charged residues provides information about the location of the protein. Intracellular proteins tend to have a higher fraction of negatively charged residues than extracellular proteins.

Total number of positively charged residues (Arg+Lys)

At neutral pH, nuclear proteins have a high relative percentage of positively charged amino acids. Nuclear proteins often bind to the negatively charged DNA, which may regulate gene expression or help to fold the DNA. Nuclear proteins often have a low percentage of aromatic residues [Andrade et al., 1998].

Amino acid distribution

Amino acids are the basic components of proteins. The amino acid distribution in a protein is simply the percentage of the different amino acids represented in a particular protein of interest. Amino acid composition is generally conserved through family-classes in different organisms which can be useful when studying a particular protein or enzymes across species borders. Another interesting observation is that amino acid composition variate slightly between

proteins from different subcellular localizations. This fact has been used in several computational methods, used for prediction of subcellular localization.

Annotation table

This table provides an overview of all the different annotations associated with the sequence and their incidence.

Dipeptide distribution

This measure is simply a count, or frequency, of all the observed adjacent pairs of amino acids (dipeptides) found in the protein. It is only possible to report neighboring amino acids. Knowledge on dipeptide composition have previously been used for prediction of subcellular localization.

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

14.5 Join sequences

CLC Protein Workbench can join several nucleotide or protein sequences into one sequence. This feature can for example be used to construct "supergenes" for phylogenetic inference by joining several disjoint genes into one. Note, that when sequences are joined, all their annotations are carried over to the new spliced sequence.

Two (or more) sequences can be joined by:

select sequences to join | **Toolbox in the Menu Bar** | **General Sequence Analyses** | **Join sequences** ( )

or **select sequences to join** | **right-click any selected sequence** | **Toolbox** | **General Sequence Analyses** | **Join sequences** ( )

This opens the dialog shown in figure 14.17.

If you have selected some sequences before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences from the selected elements. Click **Next** opens the dialog shown in figure 14.18.

In step 2 you can change the order in which the sequences will be joined. Select a sequence and use the arrows to move the selected sequence up or down.

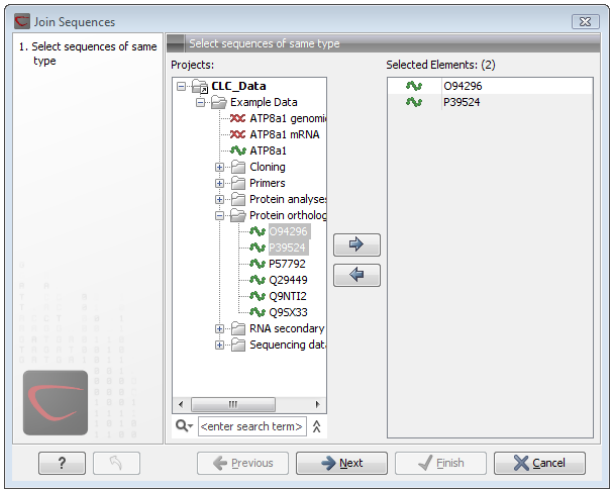


Figure 14.17: *Selecting two sequences to be joined.*

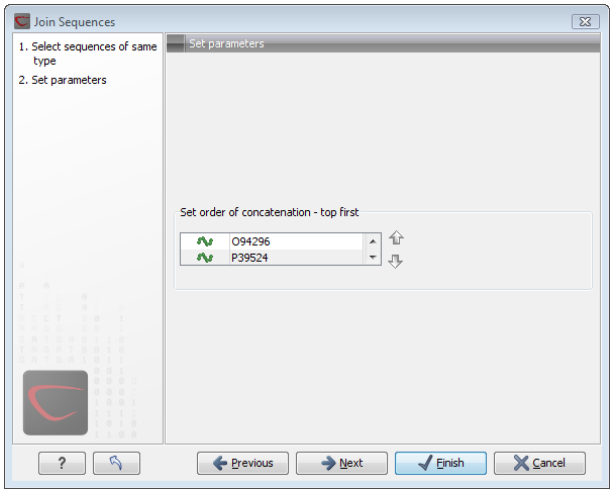


Figure 14.18: *Setting the order in which sequences are joined.*

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

The result is shown in figure 14.19.

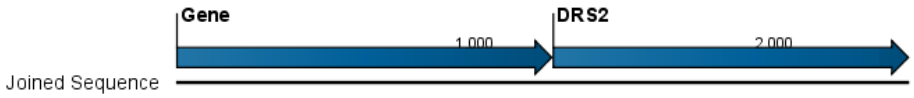


Figure 14.19: *The result of joining sequences is a new sequence containing the annotations of the joined sequences (they each had a HBB annotation).*

14.6 Pattern Discovery

With *CLC Protein Workbench* you can perform pattern discovery on both DNA and protein sequences. Advanced hidden Markov models can help to identify unknown sequence patterns across single or even multiple sequences.

In order to search for unknown patterns:

Select DNA or protein sequence(s) | **Toolbox in the Menu Bar** | **General Sequence Analyses** ( ) | **Pattern Discovery** ( )

or **right-click DNA or protein sequence(s)** | **Toolbox** | **General Sequence Analyses** ( ) | **Pattern Discovery** ( )

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

You can perform the analysis on several DNA or several protein sequences at a time. If the analysis is performed on several sequences at a time the method will search for patterns which is common between all the sequences. Annotations will be added to all the sequences and a view is opened for each sequence.

Click **Next** to adjust parameters (see figure 14.20).

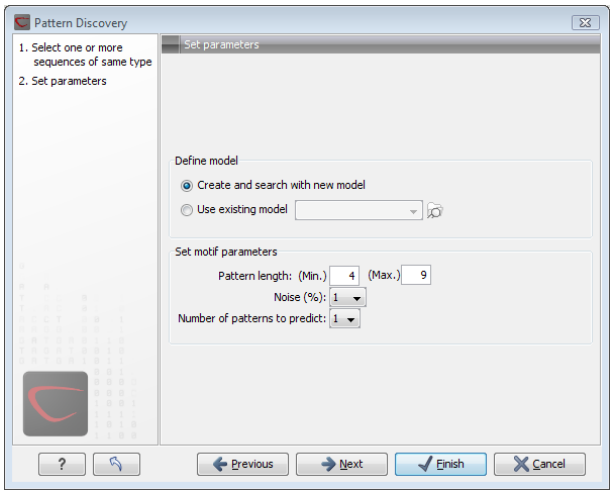


Figure 14.20: *Setting parameters for the pattern discovery. See text for details.*

In order to search unknown sequences with an already existing model:

Select to use an already existing model which is seen in figure 14.20. Models are represented with the following icon in the navigation area ( ).

14.6.1 Pattern discovery search parameters

Various parameters can be set prior to the pattern discovery. The parameters are listed below and a screen shot of the parameter settings can be seen in figure 14.20.

- **Create and search with new model.** This will create a new HMM model based on the selected sequences. The found model will be opened after the run and presented in a table view. It can be saved and used later if desired.
- **Use existing model.** It is possible to use already created models to search for the same pattern in new sequences.
- **Minimum pattern length.** Here, the minimum length of patterns to search for, can be specified.

- **Maximum pattern length.** Here, the maximum length of patterns to search for, can be specified.
- **Noise (%).** Specify noise-level of the model. This parameter has influence on the level of degeneracy of patterns in the sequence(s). The noise parameter can be 1,2,5 or 10 percent.
- **Number of different kinds of patterns to predict.** Number of iterations the algorithm goes through. After the first iteration, we force predicted pattern-positions in the first run to be member of the background: In that way, the algorithm finds new patterns in the second iteration. Patterns marked 'Pattern1' have the highest confidence. The maximal iterations to go through is 3.
- **Include background distribution.** For protein sequences it is possible to include information on the background distribution of amino acids from a range of organisms.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. This will open a view showing the patterns found as annotations on the original sequence (see figure 14.21). If you have selected several sequences, a corresponding number of views will be opened.

Figure 14.21: *Sequence view displaying two discovered patterns.*

14.6.2 Pattern search output

If the analysis is performed on several sequences at a time the method will search for patterns in the sequences and open a new view for each of the sequences, in which a pattern was discovered. Each novel pattern will be represented as an annotation of the type **Region**. More information on each found pattern is available through the tool-tip, including detailed information on the position of the pattern and quality scores.

It is also possible to get a tabular view of all found patterns in one combined table. Then each found pattern will be represented with various information on obtained scores, quality of the pattern and position in the sequence.

A table view of emission values of the actual used HMM model is presented in a table view. This model can be saved and used to search for a similar pattern in new or unknown sequences.

14.7 Motif Search

CLC Protein Workbench offers advanced and versatile options to search for known motifs represented either by a simple sequence or a more advanced regular expression. These advanced search capabilities are available for use in both DNA and protein sequences.

There are two ways to access this functionality:

- When viewing sequences, it is possible to have motifs calculated and shown on the sequence in a similar way as restriction sites (see section 17.1). This approach is called *Dynamic motifs* and is an easy way to spot known sequence motifs when working with sequences for cloning etc.
- For more refined and systematic search for motifs can be performed through the **Toolbox**. This will generate a table and optionally add annotations to the sequences.

The two approaches are described below.

14.7.1 Dynamic motifs

In the **Side Panel** of sequence views, there is a group called **Motifs** (see figure 14.22).

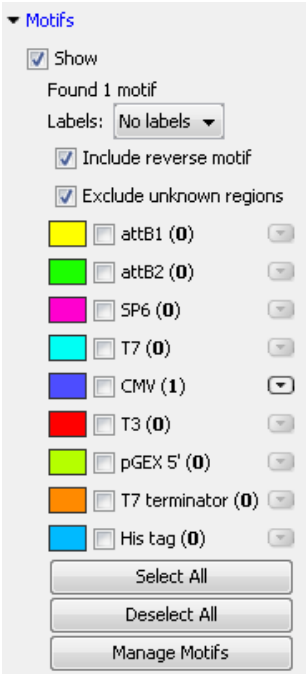


Figure 14.22: *Dynamic motifs in the Side Panel.*

The Workbench will look for the listed motifs in the sequence that is open and by clicking the check box next to the motif it will be shown in the view as illustrated in figure 14.23.

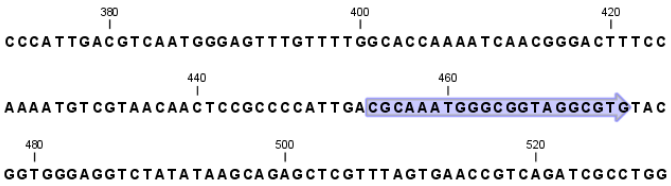


Figure 14.23: *Showing dynamic motifs on the sequence.*

This case shows the CMV promoter primer sequence which is one of the pre-defined motifs in *CLC Protein Workbench*. The motif is per default shown as a faded arrow with no text. The direction of the arrow indicates the strand of the motif.

Placing the mouse cursor on the arrow will display additional information about the motif as illustrated in figure 14.24.

Figure 14.24: *Showing dynamic motifs on the sequence.*

To add **Labels** to the motif, select the **Flag** or **Stacked** option. They will put the name of the motif as a flag above the sequence. The stacked option will stack the labels when there is more than one motif so that all labels are shown.

Below the labels option there are two options for controlling the way the sequence should be searched for motifs:

- **Include reverse motifs**. This will also find motifs on the negative strand (only available for nucleotide sequences)
- **Exclude matches in N-regions for simple motifs**. The motif search handles ambiguous characters in the way that two residues are different if they do not have any residues in common. For example: For nucleotides, *N* matches any character and *R* matches *A,G*. For proteins, *X* matches any character and *Z* matches *E,Q*. Genome sequence often have large regions with unknown sequence. These regions are very often padded with N's. Ticking this checkbox will not display hits found in N-regions and if a one residue in a motif matches to an N, it will be treated as a mismatch.

The list of motifs shown in figure 14.22 is a pre-defined list that is included with the *CLC Protein Workbench*. You can define your own set of motifs to use instead. In order to do this, you first need to create a **Motif list** ( ) (see section 14.7.4) and then click the **Manage Motifs** button. This will bring up the dialog shown in figure 14.25.

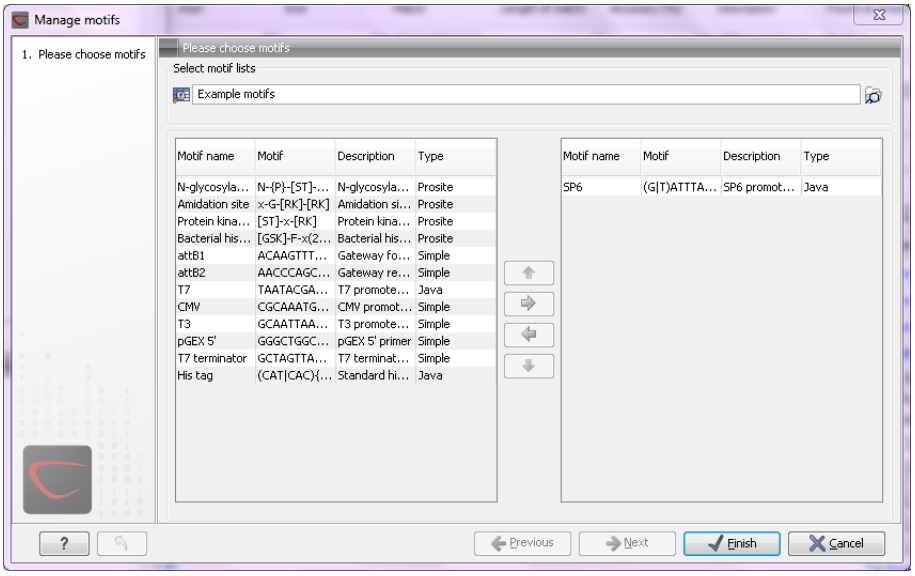


Figure 14.25: *Managing the motifs to be shown.*

At the top, select a motif list by clicking the **Browse** ( ) button. When the motif list is selected, its motifs are listed in the panel in the left-hand side of the dialog. The right-hand side panel contains the motifs that will be listed in the **Side Panel** when you click **Finish**.

14.7.2 Motif search from the Toolbox

The dynamic motifs described in section 14.7.1 provide a quick way of routinely scanning a sequence for commonly used motifs, but in some cases a more systematic approach is needed. The motif search in the **Toolbox** provides an option to search for motifs with a user-specified similarity to the target sequence, and furthermore the motifs found can be displayed in an overview table. This is particularly useful when searching for motifs on many sequences.

To start the Toolbox motif search:

Toolbox | **General Sequence Analyses ( )** | **Motif Search ( )**

Use the arrows to add or remove sequences or sequence lists from the selected elements.

You can perform the analysis on several DNA or several protein sequences at a time. If the analysis is performed on several sequences at a time the method will search for patterns in the sequences and create an overview table of the motifs found in all sequences.

Click **Next** to adjust parameters (see figure 14.26).

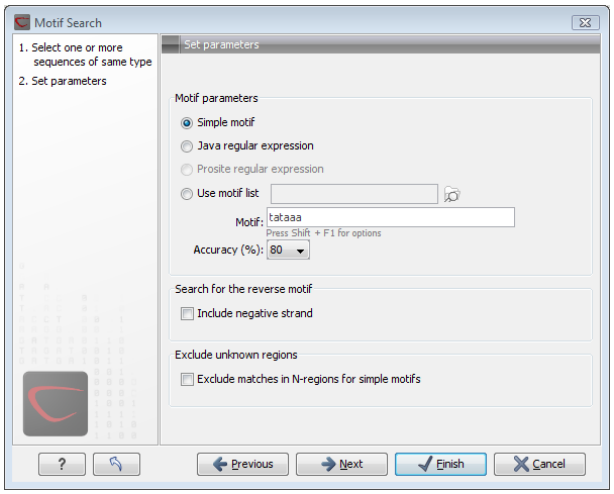


Figure 14.26: *Setting parameters for the motif search.*

The options for the motif search are:

- **Motif types.** Choose what kind of motif to be used:
  - Simple motif. Choosing this option means that you enter a simple motif, e.g. ATGATGNNATG.
  - Java regular expression. See section 14.7.3.
  - Prosite regular expression. For proteins, you can enter different protein patterns from the PROSITE database (protein patterns using regular expressions and describing specific amino acid sequences). The PROSITE database contains a great number of patterns and have been used to identify related proteins (see `http://www.expasy.org/cgi-bin/prosite-list.pl`).

  - Use motif list. Clicking the small button ( ) will allow you to select a saved motif list (see section 14.7.4).

- **Motif.** If you choose to search with a simple motif, you should enter a literal string as your motif. Ambiguous amino acids and nucleotides are allowed. Example; ATGATGNNATG. If your motif type is Java regular expression, you should enter a regular expression according to the syntax rules described in section 14.7.3. Press **Shift + F1** key for options. For proteins, you can search with a Prosite regular expression and you should enter a protein pattern from the PROSITE database.

- **Accuracy.** If you search with a simple motif, you can adjust the accuracy of the motif to the match on the sequence. If you type in a simple motif and let the accuracy be 80%, the motif search algorithm runs through the input sequence and finds all subsequences of the same length as the simple motif such that the fraction of identity between the subsequence and the simple motif is at least 80%. A motif match is added to the sequence as an annotation with the exact fraction of identity between the subsequence and the simple motif. If you use a list of motifs, the accuracy applies only to the simple motifs in the list.

- **Search for reverse motif.** This enables searching on the negative strand on nucleotide sequences.

- **Exclude unknown regions.** Genome sequence often have large regions with unknown sequence. These regions are very often padded with N's. Ticking this checkbox will not display hits found in N-regions.Motif search handles ambiguous characters in the way that two residues are different if they do not have any residues in common. For example: For nucleotides, *N* matches any character and *R* matches *A,G*. For proteins, *X* matches any character and *Z* matches *E,Q*.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. There are two types of results that can be produced:

- **Add annotations**. This will add an annotation to the sequence when a motif is found (an example is shown in figure 14.27.

- **Create table**. This will create an overview table of all the motifs found for all the input sequences.

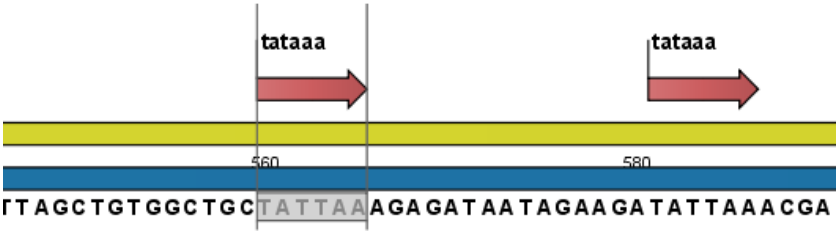


Figure 14.27: *Sequence view displaying the pattern found. The search string was 'tataaa'.*

14.7.3 Java regular expressions

A regular expressions is a string that describes or matches a set of strings, according to certain syntax rules. They are usually used to give a concise description of a set, without

having to list all elements. The simplest form of a regular expression is a literal string. The syntax used for the regular expressions is the Java regular expression syntax (see http://java.sun.com/docs/books/tutorial/essential/regex/index.html). Below is listed some of the most important syntax rules which are also shown in the help pop-up when you press Shift + F1:

[A-Z] will match the characters *A* through *Z* (Range). You can also put single characters between the brackets: The expression *[AGT]* matches the characters *A*, *G* or *T*.

[A-D[M-P]] will match the characters *A* through *D* and *M* through *P* (Union). You can also put single characters between the brackets: The expression *[AG[M-P]]* matches the characters *A*, *G* and *M* through *P*.

[A-M&&[H-P]] will match the characters between *A* and *M* lying between *H* and *P* (Intersection). You can also put single characters between the brackets. The expression *[A-M&&[HGTDA]]* matches the characters *A* through *M* which is *H*, *G*, *T*, *D* or *A*.

[^A-M] will match any character except those between *A* and *M* (Excluding). You can also put single characters between the brackets: The expression *[^AG]* matches any character except *A* and *G*.

[A-Z&&[^M-P]] will match any character *A* through *Z* except those between *M* and *P* (Subtraction). You can also put single characters between the brackets: The expression *[A-P&&[^CG]]* matches any character between *A* and *P* except *C* and *G*.

The symbol . matches any character.

X{n} will match a repetition of an element indicated by following that element with a numerical value or a numerical range between the curly brackets. For example, *ACG{2}* matches the string *ACGG* and *(ACG){2}* matches *ACGACG*.

X{n,m} will match a certain number of repetitions of an element indicated by following that element with two numerical values between the curly brackets. The first number is a lower limit on the number of repetitions and the second number is an upper limit on the number of repetitions. For example, *ACT{1,3}* matches *ACT, ACTT* and *ACTTT*.

X{n,} represents a repetition of an element at least *n* times. For example, *(AC){2,}* matches all strings *ACAC*, *ACACAC*, *ACACACAC*,...

The symbol ^ restricts the search to the beginning of your sequence. For example, if you search through a sequence with the regular expression *^AC*, the algorithm will find a match if *AC* occurs in the beginning of the sequence.

The symbol $ restricts the search to the end of your sequence. For example, if you search through a sequence with the regular expression *GT$*, the algorithm will find a match if *GT* occurs in the end of the sequence.

Examples

The expression *[ACG][^AC]G{2}* matches all strings of length *4*, where the first character is *A,C* or *G* and the second is any character except *A,C* and the third and fourth character is *G*. The expression *G.[^A]$* matches all strings of length *3* in the end of your sequence, where the first character is *C*, the second any character and the third any character except *A*.

14.7.4 Create motif list

CLC Protein Workbench offers advanced and versatile options to create lists of sequence patterns or known motifs represented either by a literal string or a regular expression.

A motif list is created from the Toolbox:

Toolbox | **General Sequence Analyses** | **Create Motif List (** **)**

This will open an empty list where you can add motifs by clicking the **Add** ( ) button at the bottom of the view. This will open a dialog shown in figure 14.28.

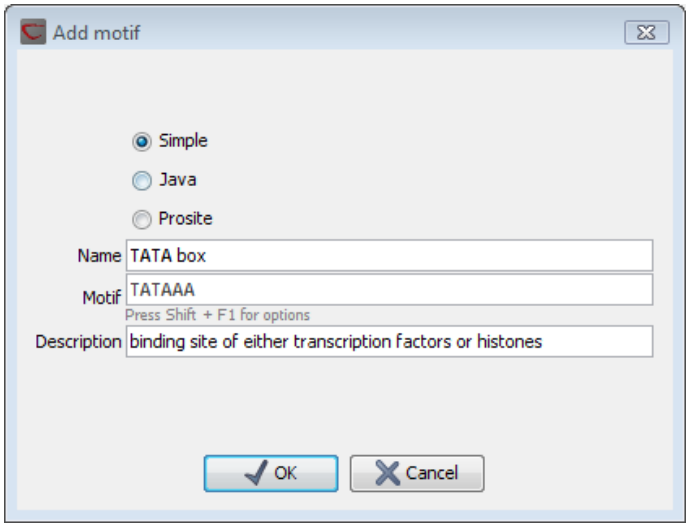


Figure 14.28: *Entering a new motif in the list.*

In this dialog, you can enter the following information:

- **Name**. The name of the motif. In the result of a motif search, this name will appear as the name of the annotation and in the result table.
- **Motif**. The actual motif. See section 14.7.2 for more information about the syntax of motifs.
- **Description**. You can enter a description of the motif. In the result of a motif search, the description will appear in the result table and added as a note to the annotation on the sequence (visible in the **Annotation table** ( ) or by placing the mouse cursor on the annotation).
- **Type**. You can enter three different types of motifs: Simple motifs, java regular expressions or PROSITE regular expression. Read more in section 14.7.2.

The motif list can contain a mix of different types of motifs. This is practical because some motifs can be described with the simple syntax, whereas others need the more advanced regular expression syntax.

Instead of manually adding motifs, you can **Import From Fasta File** ( ). This will show a dialog where you can select a fasta file on your computer and use this to create motifs. This will automatically take the name, description and sequence information from the fasta file, and put it into the motif list. The motif type will be "simple".

Besides adding new motifs, you can also edit and delete existing motifs in the list. To edit a motif, either double-click the motif in the list, or select and click the **Edit** ( ) button at the bottom of the view.

To delete a motif, select it and press the Delete key on the keyboard. Alternatively, click **Delete** ( ) in the **Tool bar**.

Save the motif list in the **Navigation Area**, and you will be able to use for Motif Search ( ) (see section 14.7).

Chapter 15

Nucleotide analyses

Contents

15.1 Convert DNA to RNA 225
15.2 Convert RNA to DNA 226
15.3 Reverse complements of sequences 227
15.4 Reverse sequence 228
15.5 Translation of DNA or RNA to protein 228
15.5.1 Translate part of a nucleotide sequence 230
15.6 Find open reading frames 230
15.6.1 Open reading frame parameters 230

CLC Protein Workbench offers different kinds of sequence analyses, which only apply to DNA and RNA.

15.1 Convert DNA to RNA

CLC Protein Workbench lets you convert a DNA sequence into RNA, substituting the T residues (Thymine) for U residues (Urasil):

select a DNA sequence in the Navigation Area | Toolbox in the Menu Bar | Nucleotide Analyses () **| Convert DNA to RNA** ()

or **right-click a sequence in Navigation Area | Toolbox | Nucleotide Analyses** () **| Convert DNA to RNA** ()

This opens the dialog displayed in figure 15.1:

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

Note! You can select multiple DNA sequences and sequence lists at a time. If the sequence list contains RNA sequences as well, they will not be converted.

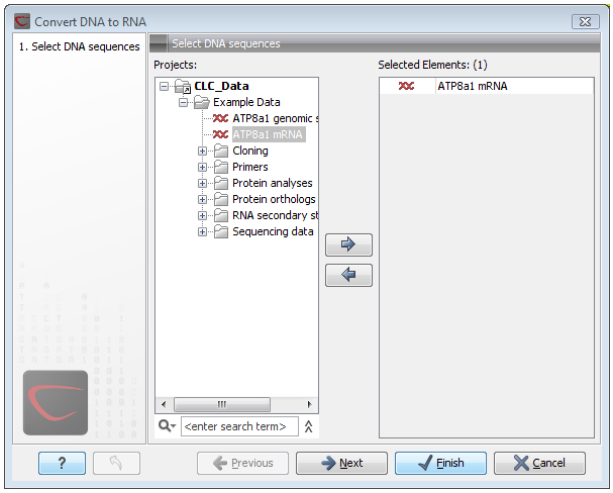


Figure 15.1: *Translating DNA to RNA.*

15.2 Convert RNA to DNA

CLC Protein Workbench lets you convert an RNA sequence into DNA, substituting the U residues (Urasil) for T residues (Thymine):

select an RNA sequence in the Navigation Area | **Toolbox in the Menu Bar** | **Nucleotide Analyses** ( ) | **Convert RNA to DNA** ( )

or **right-click a sequence in Navigation Area** | **Toolbox** | **Nucleotide Analyses** ( ) | **Convert RNA to DNA** ( )

This opens the dialog displayed in figure 15.2:

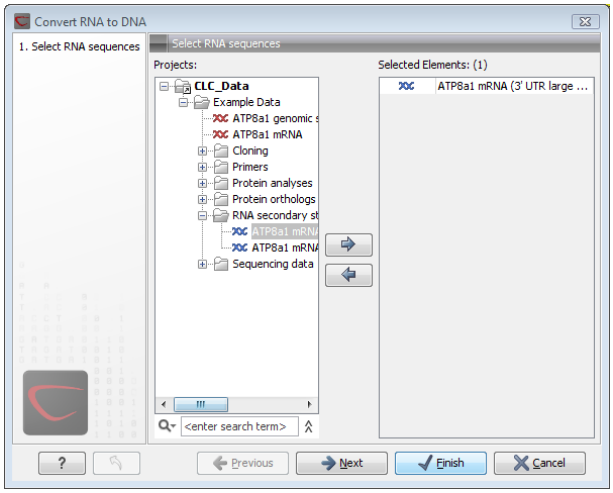


Figure 15.2: *Translating RNA to DNA.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the new DNA sequence. The new sequence is not saved automatically. To save the sequence, drag it into the **Navigation Area** or press Ctrl

+ S (⌘ + S on Mac) to activate a save dialog.

Note! You can select multiple RNA sequences and sequence lists at a time. If the sequence list contains DNA sequences as well, they will not be converted.

15.3 Reverse complements of sequences

CLC Protein Workbench is able to create the reverse complement of a nucleotide sequence. By doing that, a new sequence is created which also has all the annotations reversed since they now occupy the opposite strand of their previous location.

To quickly obtain the reverse complement of a sequence or part of a sequence, you may select a region on the negative strand and open it in a new view:

right-click a selection on the negative strand | **Open selection in New View** ( )

By doing that, the sequence will be reversed. This is only possible when the double stranded view option is enabled. It is possible to copy the selection and paste it in a word processing program or an e-mail. To obtain a reverse complement of an entire sequence:

select a sequence in the Navigation Area | **Toolbox in the Menu Bar** | **Nucleotide Analyses** ( ) | **Reverse Complement** ( )

or **right-click a sequence in Navigation Area** | **Toolbox** | **Nucleotide Analyses** ( ) | **Reverse Complement** ( )

This opens the dialog displayed in figure 15.3:

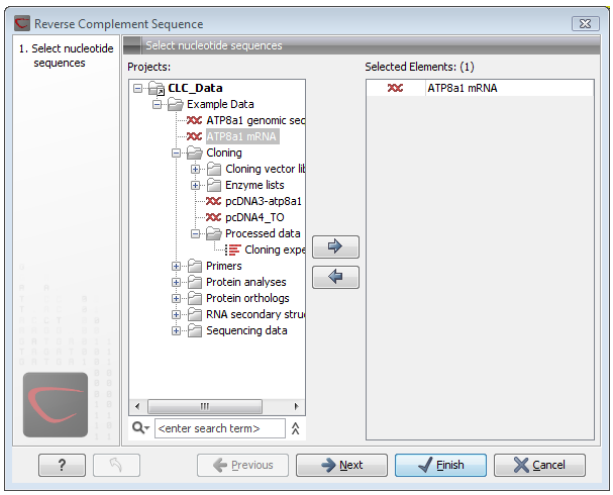


Figure 15.3: *Creating a reverse complement sequence.*

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

This will open a new view in the **View Area** displaying the reverse complement of the selected sequence. The new sequence is not saved automatically. To save the sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

15.4 Reverse sequence

CLC Protein Workbench is able to create the reverse of a nucleotide sequence. By doing that, a new sequence is created which also has all the annotations reversed since they now occupy the opposite strand of their previous location.

Note! This is not the same as a reverse complement. If you wish to create the reverse complement, please refer to section 15.3.

select a sequence in the Navigation Area | **Toolbox in the Menu Bar** | **Nucleotide Analyses** ( ) | **Reverse Sequence** ( )

This opens the dialog displayed in figure 15.4:

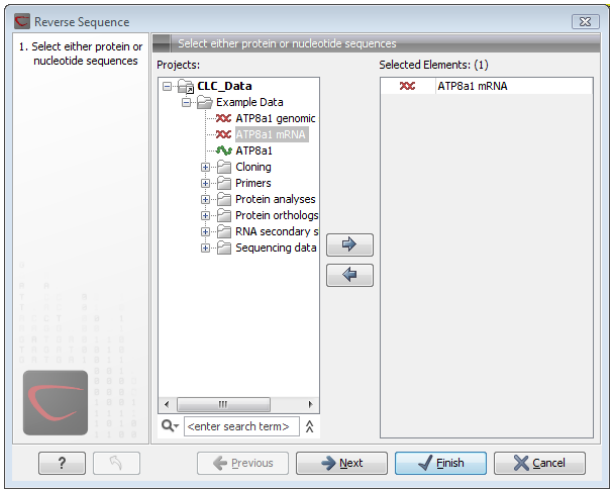


Figure 15.4: *Reversing a sequence.*

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

Note! This is not the same as a reverse complement. If you wish to create the reverse complement, please refer to section 15.3.

15.5 Translation of DNA or RNA to protein

In *CLC Protein Workbench* you can translate a nucleotide sequence into a protein sequence using the **Toolbox** tools. Usually, you use the +1 reading frame which means that the translation starts from the first nucleotide. Stop codons result in an asterisk being inserted in the protein sequence at the corresponding position. It is possible to translate in any combination of the six reading frames in one analysis. To translate:

select a nucleotide sequence | **Toolbox in the Menu Bar** | **Nucleotide Analyses** ( ) | **Translate to Protein** ( )

or **right-click a nucleotide sequence** | **Toolbox** | **Nucleotide Analyses** ( ) | **Translate to Protein** ( )

This opens the dialog displayed in figure 15.5:

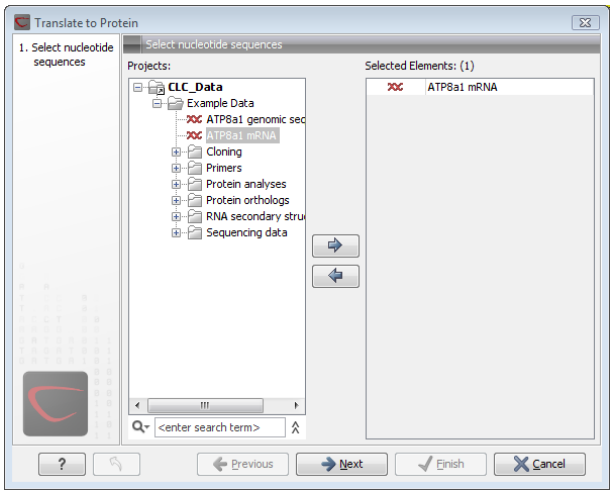


Figure 15.5: *Choosing sequences for translation.*

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

Clicking **Next** generates the dialog seen in figure 15.6:

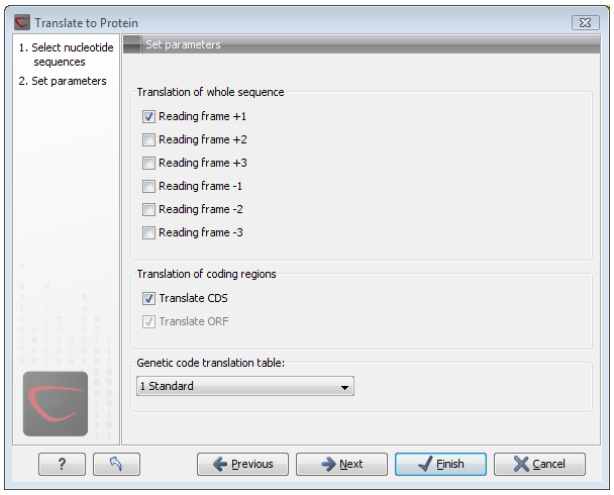


Figure 15.6: *Choosing +1 and +3 reading frames, and the standard translation table.*

Here you have the following options:

Reading frames If you wish to translate the whole sequence, you must specify the reading frame for the translation. If you select e.g. two reading frames, two protein sequences are generated.

Translate coding regions You can choose to translate regions marked by and CDS or ORF annotation. This will generate a protein sequence for each CDS or ORF annotation on the sequence.

Genetic code translation table Lets you specify the genetic code for the translation. The translation tables are occasionally updated from NCBI. The tables are not available in this

printable version of the user manual. Instead, the tables are included in the **Help**-menu in the **Menu Bar** (in the appendix).

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. The newly created protein is shown, but is not saved automatically.

To save a protein sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to activate a save dialog.

15.5.1 Translate part of a nucleotide sequence

If you want to make separate translations of *all* the coding regions of a nucleotide sequence, you can check the option: "Translate CDS and ORF" in the translation dialog (see figure 15.6).

If you want to translate a *specific* coding region, which is annotated on the sequence, use the following procedure:

Open the nucleotide sequence | **right-click the ORF or CDS annotation** | **Translate CDS/ORF ( )** | **choose a translation table** | **OK**

If the annotation contains information about the translation, this information will be used, and you do not have to specify a translation table.

The CDS and ORF annotations are colored yellow as default.

15.6 Find open reading frames

The *CLC Protein Workbench* **Find Open Reading Frames** function can be used to find all open reading frames (ORF) in a sequence, or, by choosing particular start codons to use, it can be used as a rudimentary gene finder. ORFs identified will be shown as annotations on the sequence. You have the option of choosing a translation table, the start codons to use, minimum ORF length as well as a few other parameters. These choices are explained in this section.

To find open reading frames:

select a nucleotide sequence | **Toolbox in the Menu Bar** | **Nucleotide Analyses ( )** | **Find Open Reading Frames ( )**

or **right-click a nucleotide sequence** | **Toolbox** | **Nucleotide Analyses ()** | **Find Open Reading Frames ()**

This opens the dialog displayed in figure 15.7:

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

If you want to adjust the parameters for finding open reading frames click **Next**.

15.6.1 Open reading frame parameters

This opens the dialog displayed in figure 15.8:

The adjustable parameters for the search are:

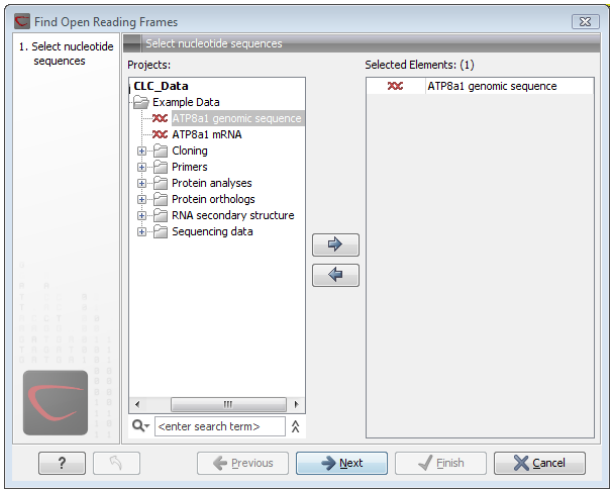


Figure 15.7: *Create Reading Frame dialog.*

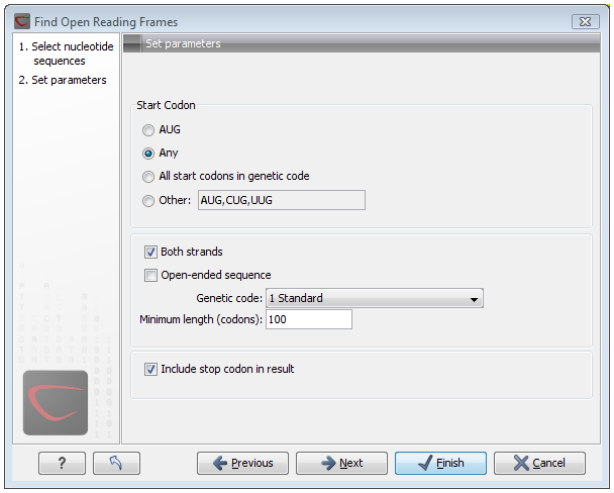


Figure 15.8: *Create Reading Frame dialog.*

- **Start codon**:
  - **AUG**. Most commonly used start codon.
  - **Any**. Find all open reading frames.
  - **All start codons in genetic code**.
  - **Other**. Here you can specify a number of start codons separated by commas.
- **Both strands**. Finds reading frames on both strands.
- **Open-ended Sequence**. Allows the ORF to start or end outside the sequence. If the sequence studied is a part of a larger sequence, it may be advantageous to allow the ORF to start or end outside the sequence.
- **Genetic code translation table**.
- **Include stop codon in result** The ORFs will be shown as annotations which can include the stop codon if this option is checked. The translation tables are occasionally updated from NCBI. The tables are not available in this printable version of the user manual. Instead, the tables are included in the **Help**-menu in the **Menu Bar** (in the appendix).

- **Minimum Length**. Specifies the minimum length for the ORFs to be found. The length is specified as number of codons.

Using open reading frames for gene finding is a fairly simple approach which is likely to predict genes which are not real. Setting a relatively high minimum length of the ORFs will reduce the number of false positive predictions, but at the same time short genes may be missed (see figure 15.9).

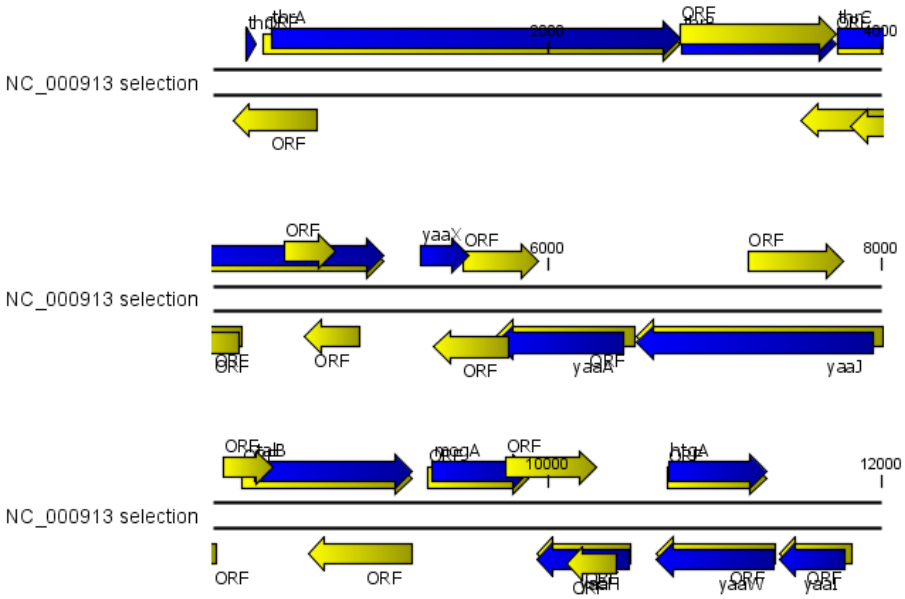


Figure 15.9: *The first 12,000 positions of the* E. coli *sequence NC_000913 downloaded from GenBank. The blue (dark) annotations are the genes while the yellow (brighter) annotations are the ORFs with a length of at least 100 amino acids. On the positive strand around position 11,000, a gene starts before the ORF. This is due to the use of the standard genetic code rather than the bacterial code. This particular gene starts with CTG, which is a start codon in bacteria. Two short genes are entirely missing, while a handful of open reading frames do not correspond to any of the annotated genes.*

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

Finding open reading frames is often a good first step in annotating sequences such as cloning vectors or bacterial genomes. For eukaryotic genes, ORF determination may not always be very helpful since the intron/exon structure is not part of the algorithm.

Chapter 16

Protein analyses

Contents

16.1 Signal peptide prediction 234
16.1.1 Signal peptide prediction parameter settings 234
16.1.2 Signal peptide prediction output 235
16.1.3 Bioinformatics explained: Prediction of signal peptides 235
16.2 Protein charge 240
16.2.1 Modifying the layout 241
16.3 Transmembrane helix prediction 241
16.4 Antigenicity 242
16.4.1 Plot of antigenicity 243
16.4.2 Antigenicity graphs along sequence 244
16.5 Hydrophobicity 244
16.5.1 Hydrophobicity plot 245
16.5.2 Hydrophobicity graphs along sequence 246
16.5.3 Bioinformatics explained: Protein hydrophobicity 247
16.6 Pfam domain search 249
16.6.1 Pfam search parameters 250
16.6.2 Download and installation of additional Pfam databases 251
16.7 Secondary structure prediction 251
16.8 Protein report 253
16.8.1 Protein report output 254
16.9 Reverse translation from protein into DNA 255
16.9.1 Reverse translation parameters 256
16.9.2 Bioinformatics explained: Reverse translation 257
16.10 Proteolytic cleavage detection 259
16.10.1 Proteolytic cleavage parameters 259
16.10.2 Bioinformatics explained: Proteolytic cleavage 262

CLC Protein Workbench offers a number of analyses of proteins as described in this chapter.

16.1 Signal peptide prediction

Signal peptides target proteins to the extracellular environment either through direct plasmamembrane translocation in prokaryotes or is routed through the Endoplasmatic Reticulum in eukaryotic cells. The signal peptide is removed from the resulting mature protein during translocation across the membrane. For prediction of signal peptides, we query SignalP [Nielsen et al., 1997, Bendtsen et al., 2004b] located at `http://www.cbs.dtu.dk/services/SignalP/`. Thus an active internet connection is required to run the signal peptide prediction. Additional information on SignalP and Center for Biological Sequence analysis (CBS) can be found at `http://www.cbs.dtu.dk` and in the original research papers [Nielsen et al., 1997, Bendtsen et al., 2004b].

In order to predict potential signal peptides of proteins, the D-score from the SignalP output is used for discrimination of signal peptide versus non-signal peptide (see section 16.1.3). This score has been shown to be the most accurate [Klee and Ellis, 2005] in an evaluation study of signal peptide predictors.

In order to use SignalP, *you need to download the SignalP plug-in* using the plug-in manager, see section 1.7.1.

When the plug-in is downloaded and installed, you can use it to predict signal peptides:

Select a protein sequence | **Toolbox in the Menu Bar** | **Protein Analyses (** **)** | **Signal Peptide Prediction (** **)**

or **right-click a protein sequence** | **Toolbox** | **Protein Analyses (** **)** | **Signal Peptide Prediction (** **)**

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

Click **Next** to set parameters for the SignalP analysis.

16.1.1 Signal peptide prediction parameter settings

It is possible to set different options prior to running the analysis (see figure 16.1). An organism type should be selected. The default is eukaryote.

- Eukaryote (default)
- Gram-negative bacteria
- Gram-positive bacteria

You can perform the analysis on several protein sequences at a time. This will add annotations to all the sequences and open a view for each sequence if a signal peptide is found. If no signal peptide is found in the sequence a dialog box will be shown.

The predictions obtained can either be shown as annotations on the sequence, listed in a table or be shown as the detailed and full text output from the SignalP method. This can be used to interpret borderline predictions:

- Add annotations to sequence

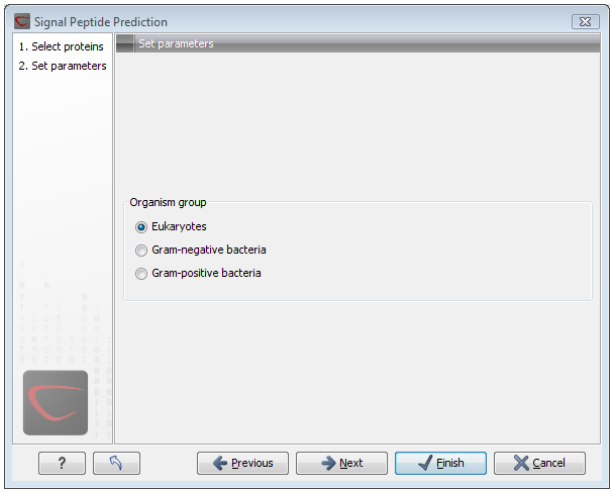


Figure 16.1: *Setting the parameters for signal peptide prediction.*

- Create table
- Text

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

16.1.2 Signal peptide prediction output

After running the prediction as described above, the protein sequence will show predicted signal peptide as annotations on the original sequence (see figure 16.2).

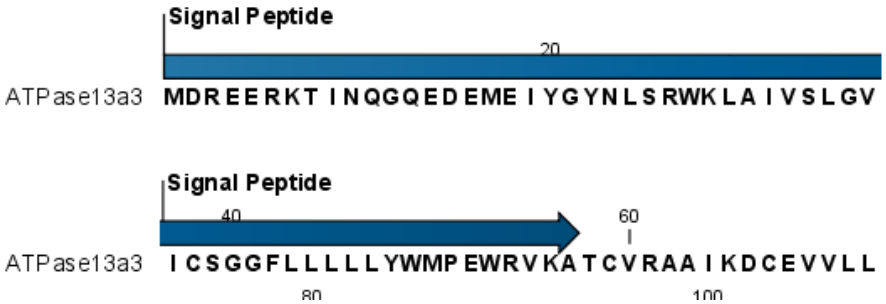


Figure 16.2: *N-terminal signal peptide shown as annotation on the sequence.*

Each annotation will carry a tooltip note saying that the corresponding annotation is predicted with SignalP version 3.0. Additional notes can be added through the **Edit annotation** ( ) right-click mouse menu. See section 10.3.2.

Undesired annotations can be removed through the **Delete Annotation ( )** right-click mouse menu. See section 10.3.4.

16.1.3 Bioinformatics explained: Prediction of signal peptides

Why the interest in signal peptides?

The importance of signal peptides was shown in 1999 when Günter Blobel received the Nobel Prize in physiology or medicine for his discovery that "proteins have intrinsic signals that govern

their transport and localization in the cell" [Blobel, 2000]. He pointed out the importance of defined peptide motifs for targeting proteins to their site of function.

Performing a query to PubMed[1] reveals that thousands of papers have been published, regarding signal peptides, secretion and subcellular localization, including knowledge of using signal peptides as vehicles for chimeric proteins for biomedical and pharmaceutical industry. Many papers describe statistical or machine learning methods for prediction of signal peptides and prediction of subcellular localization in general. After the first published method for signal peptide prediction [von Heijne, 1986], more and more methods have surfaced, although not all methods have been made available publicly.

Different types of signal peptides

Soon after Günter Blobel's initial discovery of signal peptides, more targeting signals were found. Most cell types and organisms employ several ways of targeting proteins to the extracellular environment or subcellular locations. Most of the proteins targeted for the extracellular space or subcellular locations carry specific sequence motifs (signal peptides) characterizing the type of secretion/targeting it undergoes.

Several new different signal peptides or targeting signals have been found during the later years, and papers often describe a small amino acid motif required for secretion of that particular protein. In most of the latter cases, the identified sequence motif is only found in this particular protein and as such cannot be described as a new group of signal peptides.

Describing the various types of signal peptides is beyond the scope of this text but several review papers on this topic can be found on PubMed. Targeting motifs can either be removed from, or retained in the mature protein after the protein has reached the correct and final destination. Some of the best characterized signal peptides are depicted in figure 16.3.

Numerous methods for prediction of protein targeting and signal peptides have been developed; some of them are mentioned and cited in the introduction of the SignalP research paper [Bendtsen et al., 2004b]. However, no prediction method will be able to cover all the different types of signal peptides. Most methods predicts classical signal peptides targeting to the general secretory pathway in bacteria or classical secretory pathway in eukaryotes. Furthermore, a few methods for prediction of non-classically secreted proteins have emerged [Bendtsen et al., 2004a, Bendtsen et al., 2005].

Prediction of signal peptides and subcellular localization

In the search for accurate prediction of signal peptides, many approaches have been investigated. Almost 20 years ago, the first method for prediction of classical signal peptides was published [von Heijne, 1986]. Nowadays, more sophisticated machine learning methods, such as neural networks, support vector machines, and hidden Markov models have arrived along with the increasing computational power and they all perform superior to the old weight matrix based methods [Menne et al., 2000]. Also, many other "classical" statistical approaches have been carried out, often in conjunction with machine learning methods. In the following sections, a wide range of different signal peptide and subcellular prediction methods will be described.

Most signal peptide prediction methods require the presence of the correct N-terminal end of

[1] http://www.ncbi.nlm.nih.gov/entrez/

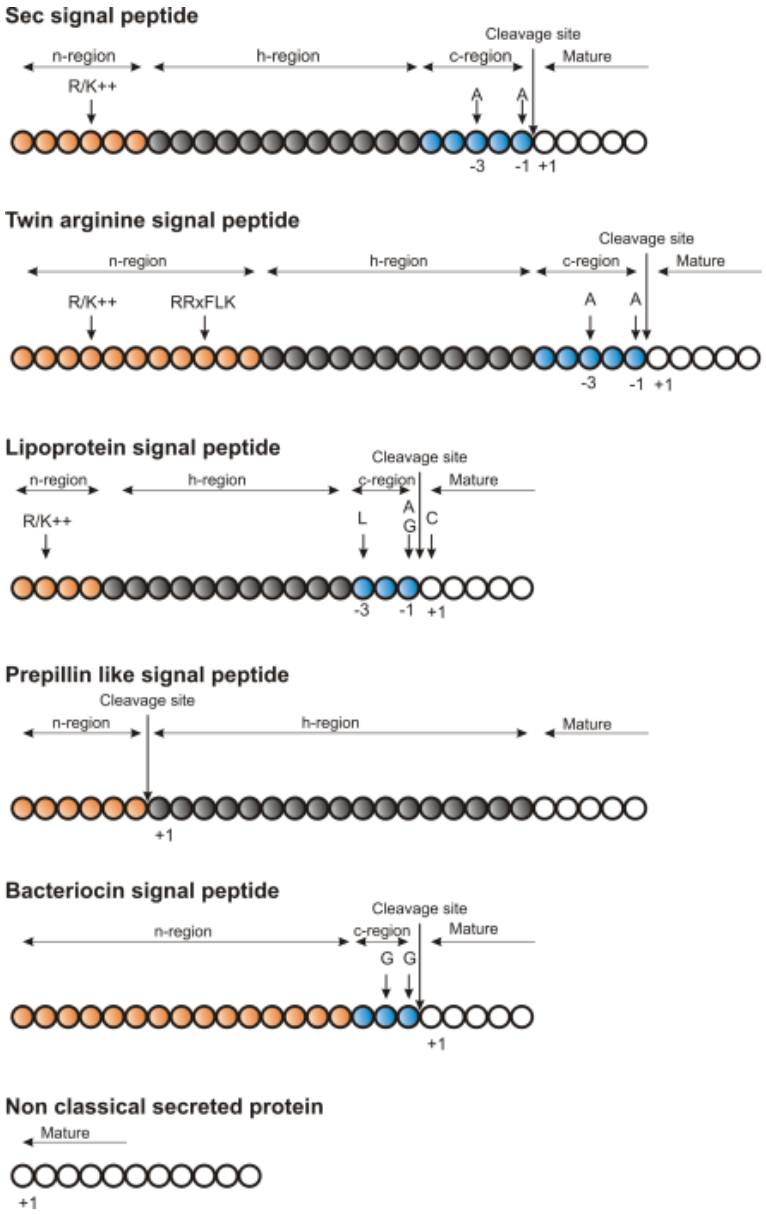


Figure 16.3: *Schematic representation of various signal peptides. Red color indicates n-region, gray color indicates h-region, cyan indicates c-region. All white circles are part of the mature protein.* $+1$ *indicates the first position of the mature protein. The length of the signal peptides is not drawn to scale.*

the preprotein for correct classification. As large scale genome sequencing projects sometimes assign the 5'-end of genes incorrectly, many proteins are annotated without the correct N-terminal [Reinhardt and Hubbard, 1998] leading to incorrect prediction of subcellular localization. These erroneous predictions can be ascribed directly to poor gene finding. Other methods for prediction of subcellular localization use information within the mature protein and therefore they are more robust to N-terminal truncation and gene finding errors.

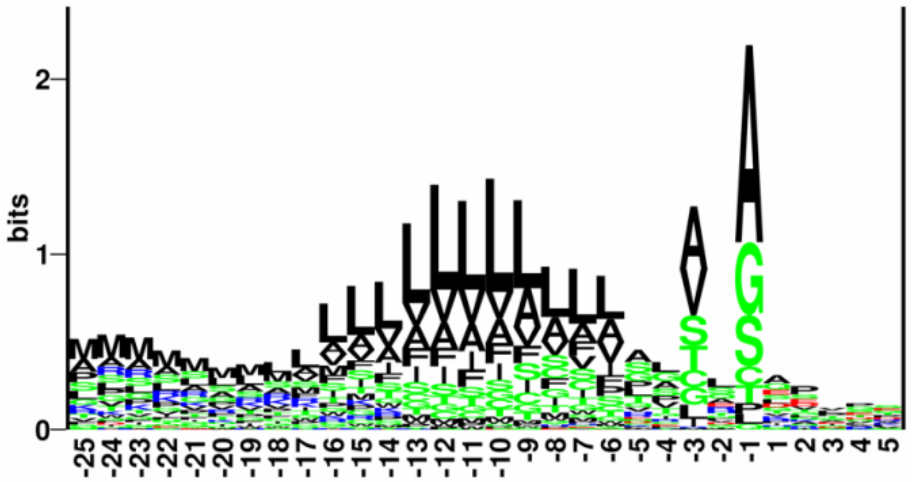


Figure 16.4: *Sequence logo of eukaryotic signal peptides, showing conservation of amino acids in bits [Schneider and Stephens, 1990]. Polar and hydrophobic residues are shown in green and black, respectively, while blue indicates positively charged residues and red negatively charged residues. The logo is based on an ungapped sequence alignment fixed at the -1 position of the signal peptides.*

The SignalP method

One of the most cited and best methods for prediction of classical signal peptides is the SignalP method [Nielsen et al., 1997, Bendtsen et al., 2004b]. In contrast to other methods, SignalP also predicts the actual cleavage site; thus the peptide which is cleaved off during translocation over the membrane. Recently, an independent research paper has rated SignalP version 3.0 to be the best standalone tool for signal peptide prediction. It was shown that the D-score which is reported by the SignalP method is the best measure for discriminating secretory from non-secretory proteins [Klee and Ellis, 2005].

SignalP is located at `http://www.cbs.dtu.dk/services/SignalP/`

What do the SignalP scores mean?

Many bioinformatics approaches or prediction tools do not give a yes/no answer. Often the user is facing an interpretation of the output, which can be either numerical or graphical. Why is that? In clear-cut examples there are no doubt; yes: this is a signal peptide! But, in borderline cases it is often convenient to have more information than just a yes/no answer. Here a graphical output can aid to interpret the correct answer. An example is shown in figure 16.5.

The graphical output from SignalP (neural network) comprises three different scores, *C*, *S* and *Y*. Two additional scores are reported in the SignalP3-NN output, namely the *S-mean* and the *D-score*, but these are only reported as numerical values.

For each organism class in SignalP; Eukaryote, Gram-negative and Gram-positive, two different neural networks are used, one for predicting the actual signal peptide and one for predicting the position of the signal peptidase I (SPase I) cleavage site. The *S-score* for the signal peptide prediction is reported for every single amino acid position in the submitted sequence, with high scores indicating that the corresponding amino acid is part of a signal peptide, and low scores indicating that the amino acid is part of a mature protein.

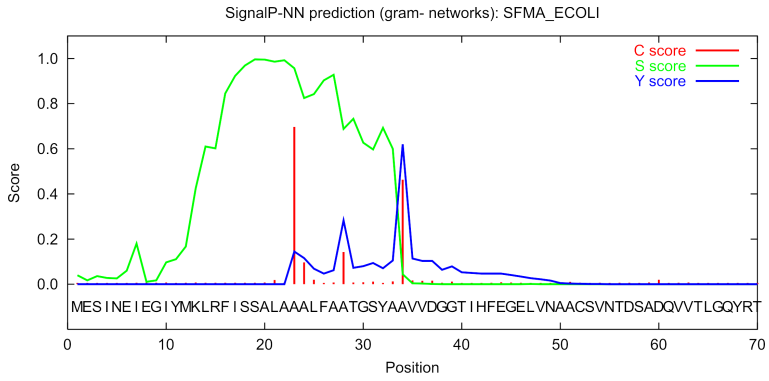


Figure 16.5: *Graphical output from the SignalP method of Swiss-Prot entry* `SFMA_ECOLI`*. Initially this seemed like a borderline prediction, but closer inspection of the sequence revealed an internal methionine at position 12, which could indicate a erroneously annotated start of the protein. Later this protein was re-annotated by Swiss-Prot to start at the M in position 12. See the text for description of the scores.*

The *C-score* is the "cleavage site" score. For each position in the submitted sequence, a C-score is reported, which should only be significantly high at the cleavage site. Confusion is often seen with the position numbering of the cleavage site. When a cleavage site position is referred to by a single number, the number indicates the first residue in the mature protein. This means that a reported cleavage site between amino acid 26-27 corresponds to the mature protein starting at (and include) position 27.

Y-max is a derivative of the C-score combined with the S-score resulting in a better cleavage site prediction than the raw C-score alone. This is due to the fact that multiple high-peaking C-scores can be found in one sequence, where only one is the true cleavage site. The cleavage site is assigned from the Y-score where the slope of the S-score is steep and a significant C-score is found.

The *S-mean* is the average of the S-score, ranging from the N-terminal amino acid to the amino acid assigned with the highest Y-max score, thus the S-mean score is calculated for the length of the predicted signal peptide. The S-mean score was in SignalP version 2.0 used as the criteria for discrimination of secretory and non-secretory proteins.

The *D-score* is introduced in SignalP version 3.0 and is a simple average of the S-mean and Y-max score. The score shows superior discrimination performance of secretory and non-secretory proteins to that of the S-mean score which was used in SignalP version 1 and 2.

For non-secretory proteins all the scores represented in the SignalP3-NN output should ideally be very low.

The hidden Markov model calculates the probability of whether the submitted sequence contains a signal peptide or not. The eukaryotic HMM model also reports the probability of a signal anchor, previously named uncleaved signal peptides. Furthermore, the cleavage site is assigned by a probability score together with scores for the n-region, h-region, and c-region of the signal peptide, if it is found.

Other useful resources

`http://www.cbs.dtu.dk/services/SignalP`

Pubmed entries for some of the original papers.

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=9051728&query_hl=1&itool=pubmed_docsum

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&list_uids=15223320&dopt=Citation

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

16.2 Protein charge

In *CLC Protein Workbench* you can create a graph in the electric charge of a protein as a function of pH. This is particularly useful for finding the net charge of the protein at a given pH. This knowledge can be used e.g. in relation to isoelectric focusing on the first dimension of 2D-gel electrophoresis. The isoelectric point (pI) is found where the net charge of the protein is zero. The calculation of the protein charge does not include knowledge about any potential post-translational modifications the protein may have.

The pKa values reported in the literature may differ slightly, thus resulting in different looking graphs of the protein charge plot compared to other programs.

In order to calculate the protein charge:

Select a protein sequence | **Toolbox in the Menu Bar** | **Protein Analyses** ( ) | **Create Protein Charge Plot** ( )

or **right-click a protein sequence** | **Toolbox** | **Protein Analyses** () | **Create Protein Charge Plot** ( )

This opens the dialog displayed in figure 16.6:

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

You can perform the analysis on several protein sequences at a time. This will result in one output graph showing protein charge graphs for the individual proteins.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

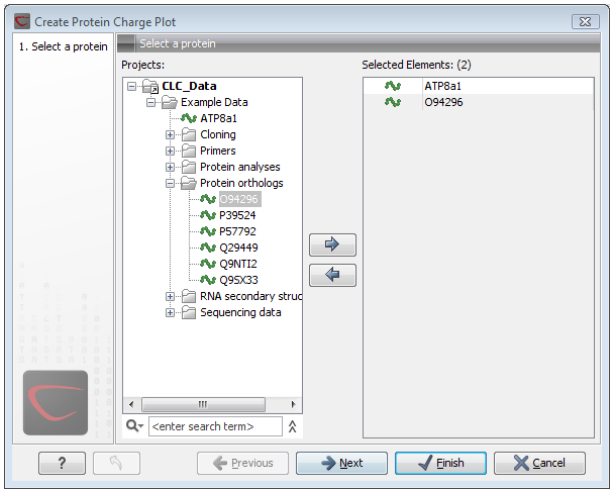


Figure 16.6: *Choosing protein sequences to calculate protein charge.*

16.2.1 Modifying the layout

Figure 16.7 shows the electrical charges for three proteins. In the **Side Panel** to the right, you can modify the layout of the graph.

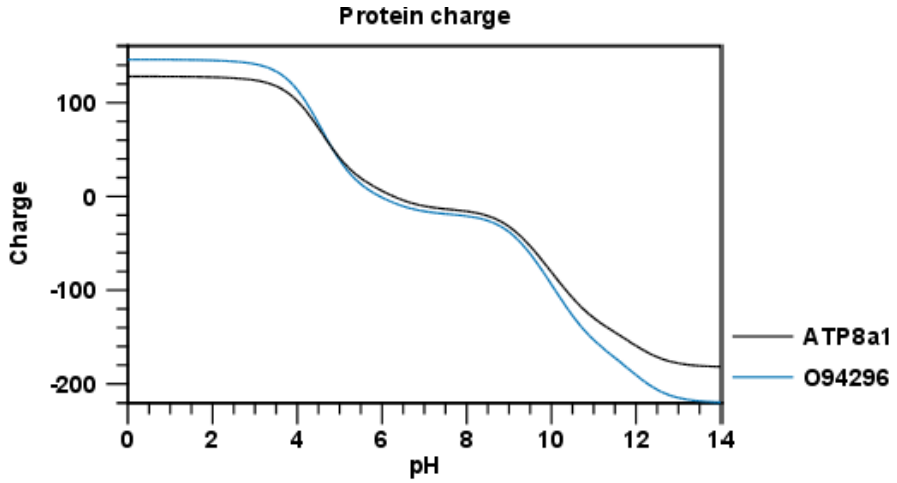


Figure 16.7: *View of the protein charge.*

See section B in the appendix for information about the graph view.

16.3 Transmembrane helix prediction

Many proteins are integral membrane proteins. Most membrane proteins have hydrophobic regions which span the hydrophobic core of the membrane bi-layer and hydrophilic regions located on the outside or the inside of the membrane. Many receptor proteins have several transmembrane helices spanning the cellular membrane.

For prediction of transmembrane helices, *CLC Protein Workbench* uses TMHMM version 2.0 [Krogh et al., 2001] located at `http://www.cbs.dtu.dk/services/TMHMM/`, thus an active internet connection is required to run the transmembrane helix prediction. Additional information on THMHH and Center for Biological Sequence analysis (CBS) can be found at

http://www.cbs.dtu.dk and in the original research paper [Krogh et al., 2001].

In order to use the transmembrane helix prediction, *you need to download the plug-in* using the plug-in manager (see section 1.7.1).

When the plug-in is downloaded and installed, you can use it to predict transmembrane helices:

Select a protein sequence | **Toolbox in the Menu Bar** | **Protein Analyses ( )** | **Transmembrane Helix Prediction ( )**

or **right-click a protein sequence** | **Toolbox** | **Protein Analyses ( )** | **Transmembrane Helix Prediction ( )**

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

The predictions obtained can either be shown as annotations on the sequence, in a table or as the detailed and text output from the TMHMM method.

- Add annotations to sequence
- Create table
- Text

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

You can perform the analysis on several protein sequences at a time. This will add annotations to all the sequences and open a view for each sequence if a transmembrane helix is found. If a transmembrane helix is not found a dialog box will be presented.

After running the prediction as described above, the protein sequence will show predicted transmembrane helices as annotations on the original sequence (see figure 16.8). Moreover, annotations showing the topology will be shown. That is, which part the proteins is located on the inside or on the outside.

Each annotation will carry a tooltip note saying that the corresponding annotation is predicted with TMHMM version 2.0. Additional notes can be added through the **Edit annotation** ( ) right-click mouse menu. See section 10.3.2.

Undesired annotations can be removed through the **Delete Annotation ( )** right-click mouse menu. See section 10.3.4.

16.4 Antigenicity

CLC Protein Workbench can help to identify antigenic regions in protein sequences in different ways, using different algorithms. The algorithms provided in the Workbench, merely plot an index of antigenicity over the sequence.

Two different methods are available.

[Welling et al., 1985] Welling et al. used information on the relative occurrence of amino acids in antigenic regions to make a scale which is useful for prediction of antigenic regions. This method

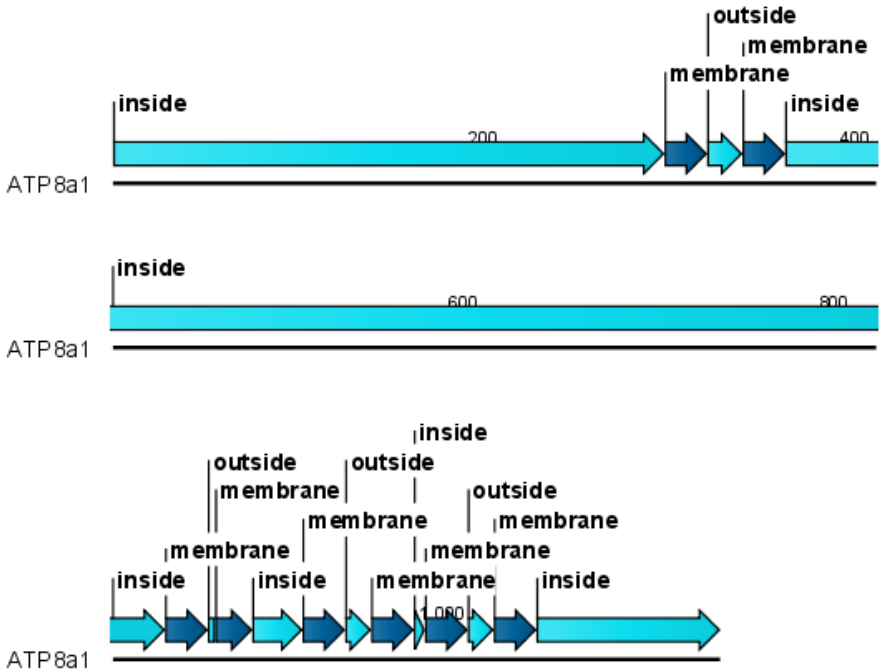


Figure 16.8: *Transmembrane segments shown as annotation on the sequence and the topology.*

is better than the Hopp-Woods scale of hydrophobicity which is also used to identify antigenic regions.

A semi-empirical method for prediction of antigenic regions has been developed [Kolaskar and Tongaonkar, 1990]. This method also includes information of surface accessibility and flexibility and at the time of publication the method was able to predict antigenic determinants with an accuracy of 75%.

Note! Similar results from the two method can not always be expected as the two methods are based on different training sets.

16.4.1 Plot of antigenicity

Displaying the antigenicity for a protein sequence in a plot is done in the following way:

> **select a protein sequence in Navigation Area | Toolbox in the Menu Bar | Protein Analyses ( ) | Create Antigenicity Plot ( )**

This opens a dialog. The first step allows you to add or remove sequences. Clicking **Next** takes you through to **Step 2**, which is displayed in figure 16.9.

The **Window size** is the width of the window where, the antigenicity is calculated. The wider the window, the less volatile the graph. You can chose from a number of antigenicity scales. Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. The result can be seen in figure 16.10.

See section B in the appendix for information about the graph view.

The level of antigenicity is calculated on the basis of the different scales. The different scales add different values to each type of amino acid. The antigenicity score is then calculated as the

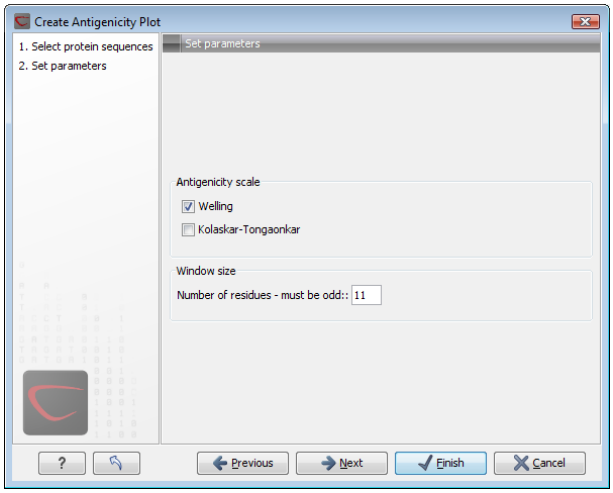


Figure 16.9: *Step two in the Antigenicity Plot allows you to choose different antigenicity scales and the window size.*

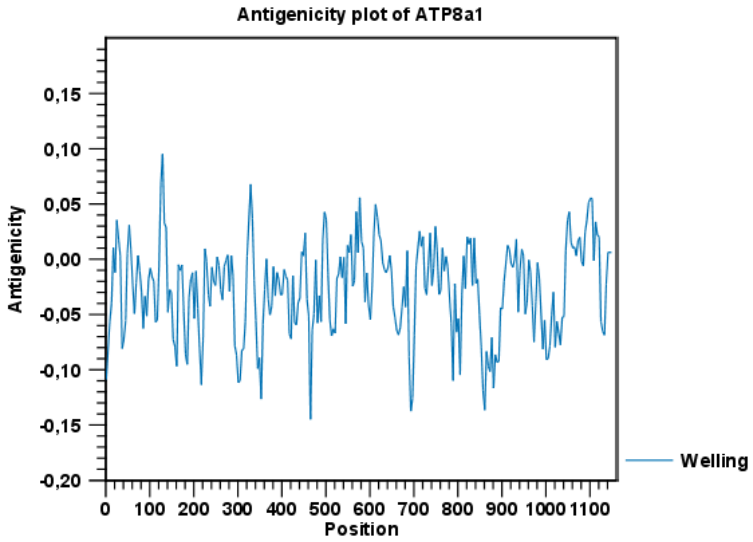


Figure 16.10: *The result of the antigenicity plot calculation and the associated Side Panel.*

sum of the values in a 'window', which is a particular range of the sequence. The window length can be set from 5 to 25 residues. The wider the window, the less fluctuations in the antigenicity scores.

16.4.2 Antigenicity graphs along sequence

Antigenicity graphs along the sequence can be displayed using the **Side Panel**. The functionality is similar to hydrophobicity (see section 16.5.2).

16.5 Hydrophobicity

CLC Protein Workbench can calculate the hydrophobicity of protein sequences in different ways, using different algorithms. (See section 16.5.3). Furthermore, hydrophobicity of sequences can be displayed as hydrophobicity plots and as graphs along sequences. In addition, *CLC Protein Workbench* can calculate hydrophobicity for several sequences at the same time, and for

alignments.

16.5.1 Hydrophobicity plot

Displaying the hydrophobicity for a protein sequence in a plot is done in the following way:

> **select a protein sequence in Navigation Area | Toolbox in the Menu Bar | Protein Analyses ( ) | Create Hydrophobicity Plot ( )**

This opens a dialog. The first step allows you to add or remove sequences. Clicking **Next** takes you through to **Step 2**, which is displayed in figure 16.11.

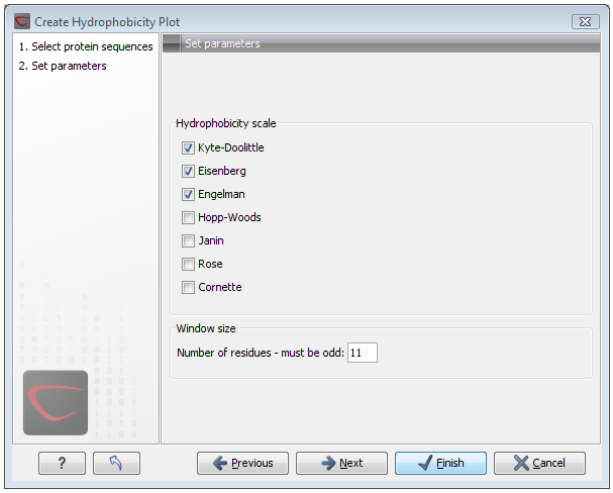


Figure 16.11: *Step two in the Hydrophobicity Plot allows you to choose hydrophobicity scale and the window size.*

The **Window size** is the width of the window where the hydrophobicity is calculated. The wider the window, the less volatile the graph. You can chose from a number of hydrophobicity scales which are further explained in section 16.5.3 Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. The result can be seen in figure 16.12.

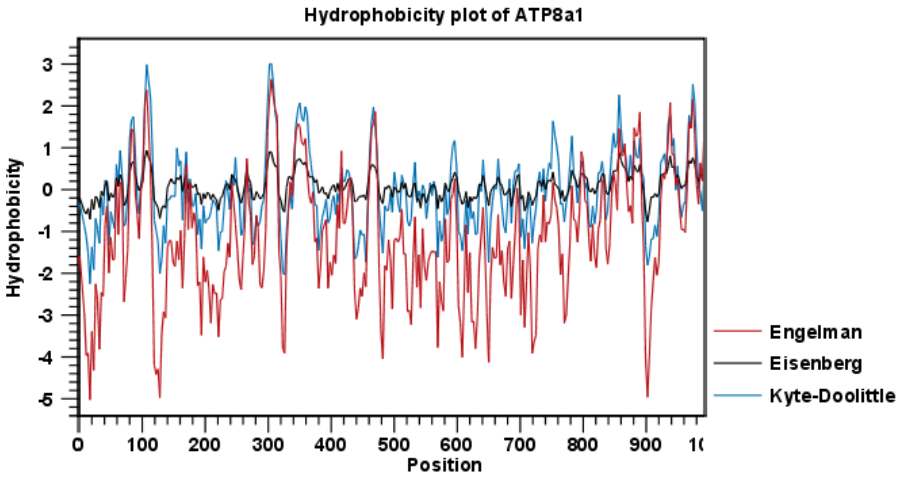


Figure 16.12: *The result of the hydrophobicity plot calculation and the associated Side Panel.*

See section B in the appendix for information about the graph view.

16.5.2 Hydrophobicity graphs along sequence

Hydrophobicity graphs along sequence can be displayed easily by activating the calculations from the **Side Panel** for a sequence.

right-click protein sequence in Navigation Area | Show | Sequence | open Protein info in Side Panel

or **double-click protein sequence in Navigation Area | Show | Sequence | open Protein info in Side Panel**

These actions result in the view displayed in figure 16.13.

Figure 16.13: *The different available scales in Protein info in* **CLC Protein Workbench**.

The level of hydrophobicity is calculated on the basis of the different scales. The different scales add different values to each type of amino acid. The hydrophobicity score is then calculated as the sum of the values in a 'window', which is a particular range of the sequence. The window length can be set from 5 to 25 residues. The wider the window, the less fluctuations in the hydrophobicity scores. (For more about the theory behind hydrophobicity, see 16.5.3).

In the following we will focus on the different ways that *CLC Protein Workbench* offers to display the hydrophobicity scores. We use Kyte-Doolittle to explain the display of the scores, but the different options are the same for all the scales. Initially there are three options for displaying the hydrophobicity scores. You can choose one, two or all three options by selecting the boxes. (See figure 16.14).

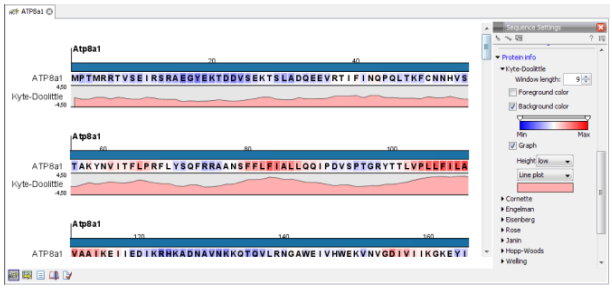


Figure 16.14: *The different ways of displaying the hydrophobicity scores, using the Kyte-Doolittle scale.*

Coloring the letters and their background. When choosing coloring of letters or coloring of their background, the color red is used to indicate high scores of hydrophobicity. A 'color-slider'

allows you to amplify the scores, thereby emphasizing areas with high (or low, blue) levels of hydrophobicity. The color settings mentioned are default settings. By clicking the color bar just below the color slider you get the option of changing color settings.

Graphs along sequences. When selecting graphs, you choose to display the hydrophobicity scores underneath the sequence. This can be done either by a line-plot or bar-plot, or by coloring. The latter option offers you the same possibilities of amplifying the scores as applies for coloring of letters. The different ways to display the scores when choosing 'graphs' are displayed in figure 16.14. Notice that you can choose the height of the graphs underneath the sequence.

16.5.3 Bioinformatics explained: Protein hydrophobicity

Calculation of hydrophobicity is important to the identification of various protein features. This can be membrane spanning regions, antigenic sites, exposed loops or buried residues. Usually, these calculations are shown as a plot along the protein sequence, making it easy to identify the location of potential protein features.

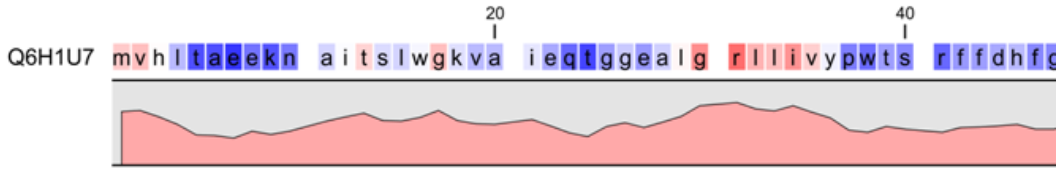


Figure 16.15: *Plot of hydrophobicity along the amino acid sequence. Hydrophobic regions on the sequence have higher numbers according to the graph below the sequence, furthermore hydrophobic regions are colored on the sequence. Red indicates regions with high hydrophobicity and blue indicates regions with low hydrophobicity.*

The hydrophobicity is calculated by sliding a fixed size window (of an odd number) over the protein sequence. At the central position of the window, the average hydrophobicity of the entire window is plotted (see figure 16.15).

Hydrophobicity scales

Several hydrophobicity scales have been published for various uses. Many of the commonly used hydrophobicity scales are described below.

Kyte-Doolittle scale. The Kyte-Doolittle scale is widely used for detecting hydrophobic regions in proteins. Regions with a positive value are hydrophobic. This scale can be used for identifying both surface-exposed regions as well as transmembrane regions, depending on the window size used. Short window sizes of 5-7 generally work well for predicting putative surface-exposed regions. Large window sizes of 19-21 are well suited for finding transmembrane domains if the values calculated are above 1.6 [Kyte and Doolittle, 1982]. These values should be used as a rule of thumb and deviations from the rule may occur.

Engelman scale. The Engelman hydrophobicity scale, also known as the GES-scale, is another scale which can be used for prediction of protein hydrophobicity [Engelman et al., 1986]. As the Kyte-Doolittle scale, this scale is useful for predicting transmembrane regions in proteins.

Eisenberg scale. The Eisenberg scale is a normalized consensus hydrophobicity scale which

shares many features with the other hydrophobicity scales [Eisenberg et al., 1984].

Hopp-Woods scale. Hopp and Woods developed their hydrophobicity scale for identification of potentially antigenic sites in proteins. This scale is basically a hydrophilic index where apolar residues have been assigned negative values. Antigenic sites are likely to be predicted when using a window size of 7 [Hopp and Woods, 1983].

Cornette scale. Cornette *et al.* computed an optimal hydrophobicity scale based on 28 published scales [Cornette et al., 1987]. This optimized scale is also suitable for prediction of alpha-helices in proteins.

Rose scale. The hydrophobicity scale by Rose *et al.* is correlated to the average area of buried amino acids in globular proteins [Rose et al., 1985]. This results in a scale which is not showing the helices of a protein, but rather the surface accessibility.

Janin scale. This scale also provides information about the accessible and buried amino acid residues of globular proteins [Janin, 1979].

Welling scale. Welling et al. used information on the relative occurrence of amino acids in antigenic regions to make a scale which is useful for prediction of antigenic regions. This method is better than the Hopp-Woods scale of hydrophobicity which is also used to identify antigenic regions.

Kolaskar-Tongaonkar. A semi-empirical method for prediction of antigenic regions has been developed [Kolaskar and Tongaonkar, 1990]. This method also includes information of surface accessibility and flexibility and at the time of publication the method was able to predict antigenic determinants with an accuracy of 75%.

Surface Probability. Display of surface probability based on the algorithm by [Emini et al., 1985]. This algorithm has been used to identify antigenic determinants on the surface of proteins.

Chain Flexibility. isplay of backbone chain flexibility based on the algorithm by [Karplus and Schulz, 1985]. It is known that chain flexibility is an indication of a putative antigenic determinant.

Many more scales have been published throughout the last three decades. Even though more advanced methods have been developed for prediction of membrane spanning regions, the simple and very fast calculations are still highly used.

Other useful resources

AAindex: Amino acid index database
http://www.genome.ad.jp/dbget/aaindex.html

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

aa	aa	Kyte-Doolittle	Hopp-Woods	Cornette	Eisenberg	Rose	Janin	Engelman (GES)
A	Alanine	1.80	-0.50	0.20	0.62	0.74	0.30	1.60
C	Cysteine	2.50	-1.00	4.10	0.29	0.91	0.90	2.00
D	Aspartic acid	-3.50	3.00	-3.10	-0.90	0.62	-0.60	-9.20
E	Glutamic acid	-3.50	3.00	-1.80	-0.74	0.62	-0.70	-8.20
F	Phenylalanine	2.80	-2.50	4.40	1.19	0.88	0.50	3.70
G	Glycine	-0.40	0.00	0.00	0.48	0.72	0.30	1.00
H	Histidine	-3.20	-0.50	0.50	-0.40	0.78	-0.10	-3.00
I	Isoleucine	4.50	-1.80	4.80	1.38	0.88	0.70	3.10
K	Lysine	-3.90	3.00	-3.10	-1.50	0.52	-1.80	-8.80
L	Leucine	3.80	-1.80	5.70	1.06	0.85	0.50	2.80
M	Methionine	1.90	-1.30	4.20	0.64	0.85	0.40	3.40
N	Asparagine	-3.50	0.20	-0.50	-0.78	0.63	-0.50	-4.80
P	Proline	-1.60	0.00	-2.20	0.12	0.64	-0.30	-0.20
Q	Glutamine	-3.50	0.20	-2.80	-0.85	0.62	-0.70	-4.10
R	Arginine	-4.50	3.00	1.40	-2.53	0.64	-1.40	-12.3
S	Serine	-0.80	0.30	-0.50	-0.18	0.66	-0.10	0.60
T	Threonine	-0.70	-0.40	-1.90	-0.05	0.70	-0.20	1.20
V	Valine	4.20	-1.50	4.70	1.08	0.86	0.60	2.60
W	Tryptophan	-0.90	-3.40	1.00	0.81	0.85	0.30	1.90
Y	Tyrosine	-1.30	-2.30	3.20	0.26	0.76	-0.40	-0.70

Table 16.1: *Hydrophobicity scales. This table shows seven different hydrophobicity scales which are generally used for prediction of e.g. transmembrane regions and antigenicity.*

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

16.6 Pfam domain search

With *CLC Protein Workbench* you can perform a search for Pfam domains on protein sequences. The Pfam database at http://pfam.sanger.ac.uk/ is a large collection of multiple sequence alignments that covers approximately 9318 protein domains and protein families [Bateman et al., 2004]. Based on the individual domain alignments, profile HMMs have been developed. These profile HMMs can be used to search for domains in unknown sequences.

Many proteins have a unique combination of domains which can be responsible, for instance, for the catalytic activities of enzymes. Pfam was initially developed to aid the annotation of the *C. elegans* genome. Annotating unknown sequences based on pairwise alignment methods by simply transferring annotation from a known protein to the unknown partner does not take domain organization into account [Galperin and Koonin, 1998]. An unknown protein may be annotated wrongly, for instance, as an enzyme if the pairwise alignment only finds a regulatory domain.

Using the Pfam search option in *CLC Protein Workbench*, you can search for domains in sequence data which otherwise do not carry any annotation information. The Pfam search option adds all found domains onto the protein sequence which was used for the search. If domains of no relevance are found they can easily be removed as described in section 10.3.4. Setting a lower cutoff value will result in fewer domains.

In *CLC Protein Workbench* we have implemented our own HMM algorithm for prediction of the Pfam domains. Thus, we do not use the original HMM implementation, HMMER http://hmmer.wustl.edu/ for domain prediction. We find the most probable state path/alignment through each profile HMM by the Viterbi algorithm and based on that we derive a new null model by averaging over the emission distributions of all *M* and *I* states that appear in the state path (*M* is a match state and *I* is an insert state). From that model we now arrive at an additive correction to the original bit-score, like it is done in the original HMMER algorithm.

In order to conduct the Pfam search:

Select a protein sequence | **Toolbox in the Menu Bar** | **Protein Analyses** ( ) | **Pfam Domain Search** ( )

or **right-click a protein sequence** | **Toolbox** | **Protein Analyses** ( ) | **Pfam Domain Search** ( )

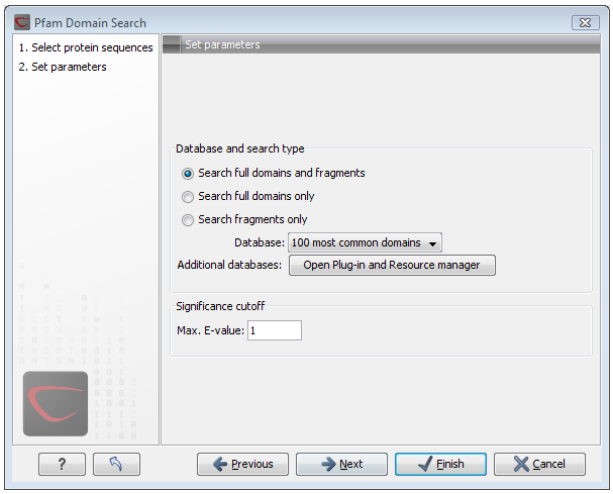


Figure 16.16: *Setting parameters for Pfam domain search.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

You can perform the analysis on several protein sequences at a time. This will add annotations to all the sequences and open a view for each sequence. Click **Next** to adjust parameters (see figure 16.16).

16.6.1 Pfam search parameters

- **Choose database and search type**
  When searching for Pfam domains it is possible to choose different databases and specify the search for full domains or fragments of domains. Only the 100 most frequent domains are included as default in *CLC Protein Workbench*. Additional databases can be downloaded directly from CLC bio's web-site at http://www.clcbio.com/resources.
  - **Search full domains and fragments.** This option allows you to search both for full domain but also for partial domains. This could be the case if a domain extends beyond the ends of a sequence
  - **Search full domains only.** Selecting this option only allows searches for full domains.

  - **Search fragments only.** Only partial domains will be found.
  - **Database.** Only the 100 most frequent domains are included as default in *CLC Protein Workbench*, but additional databases can be downloaded and installed as described in section 16.6.2.

- **Set significance cutoff.** The E-value (expectation value) is the number of hits that would be expected to have a score equal to or better than this value, by chance alone. This means that a good E-value which gives a confident prediction is much less than 1. E-values around 1 is what is expected by chance. Thus, the lower the E-value, the more specific the search for domains will be. Only positive numbers are allowed.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. This will open a view showing the found domains as annotations on the original sequence (see figure 16.17). If you have selected several sequences, a corresponding number of views will be opened.

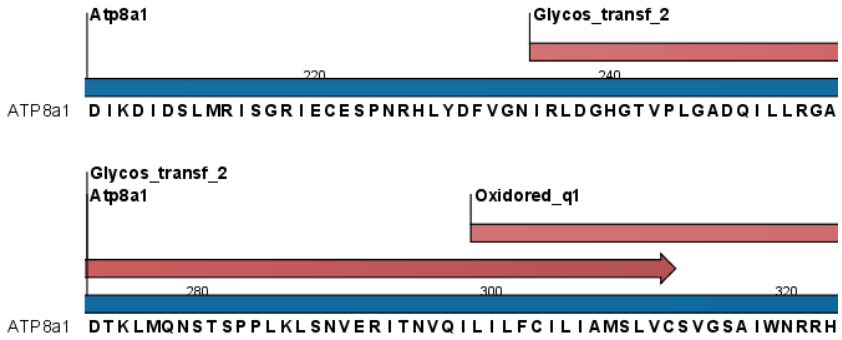

Atp8a1
Glycos_transf_2
220
240
ATP8a1 D I K D I D S L M R I S G R I E C E S P N R H L Y D F V G N I R L D G H G T V P L G A D Q I L L R G A
Glycos_transf_2
Atp8a1
Oxidored_q1
280
300
320
ATP8a1 D T K L M Q N S T S P P L K L S N V E R I T N V Q I L I L F C I L I A M S L V C S V G S A I W N R R H

Figure 16.17: *Domains annotations based on Pfam.*

Each found domain will be represented as an annotation of the type **Region**. More information on each found domain is available through the tooltip, including detailed information on the identity score which is the basis for the prediction.

For a more detailed description of the provided scores through the tool tip look at `http://pfam.sanger.ac.uk/help#tabview=tab5`.

16.6.2 Download and installation of additional Pfam databases

Additional databases can be downloaded as a resource using the **Plug-in manager** ( ) (see section 1.7.4).

If you are not able to download directly from the Plug-in manager,

16.7 Secondary structure prediction

An important issue when trying to understand protein function is to know the actual structure of the protein. Many questions that are raised by molecular biologists are directly targeted at protein structure. The alpha-helix forms a coiled rodlike structure whereas a beta-sheet show an extended sheet-like structure. Some proteins are almost devoid of alpha-helices such as

chymotrypsin (PDB_ID: 1AB9) whereas others like myoglobin (PDB_ID: 101M) have a very high content of alpha-helices.

With *CLC Protein Workbench* one can predict the secondary structure of proteins very fast. Predicted elements are alpha-helix, beta-sheet (same as beta-strand) and other regions.

Based on extracted protein sequences from the protein databank (http://www.rcsb.org/pdb/) a hidden Makov model (HMM) was trained and evaluated for performance. Machine learning methods have shown superior when it comes to prediction of secondary structure of proteins [Rost, 2001]. By far the most common structures are Alpha-helices and beta-sheets which can be predicted, and predicted structures are automatically added to the query as annotation which later can be edited.

In order to predict the secondary structure of proteins:

Select a protein sequence | **Toolbox in the Menu Bar** | **Protein Analyses ( )** | **Predict secondary structure ( )**

or **right-click a protein sequence** | **Toolbox** | **Protein Analyses ( )** | **Predict secondary structure ( )**

This opens the dialog displayed in figure 16.18:

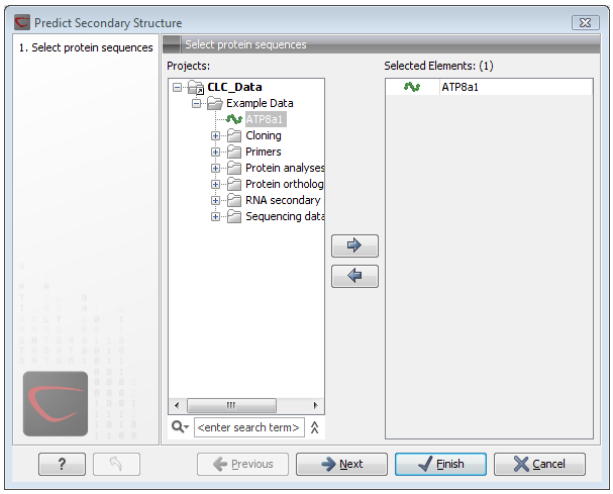


Figure 16.18: *Choosing one or more protein sequences for secondary structure prediction.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

You can perform the analysis on several protein sequences at a time. This will add annotations to all the sequences and open a view for each sequence.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

After running the prediction as described above, the protein sequence will show predicted alpha-helices and beta-sheets as annotations on the original sequence (see figure 16.19).

Each annotation will carry a tooltip note saying that the corresponding annotation is predicted with *CLC Protein Workbench*. Additional notes can be added through the **Edit Annotation ( )** right-click mouse menu. See section 10.3.2.

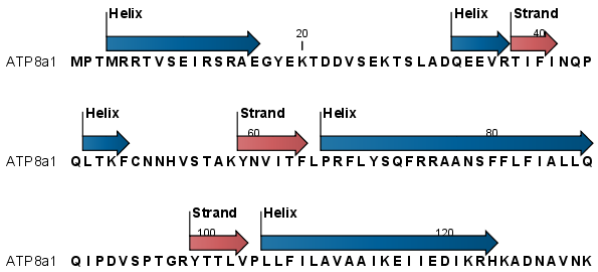


Figure 16.19: *Alpha-helices and beta-strands shown as annotations on the sequence.*

Undesired alpha-helices or beta-sheets can be removed through the **Delete Annotation** ( ) right-click mouse menu. See section 10.3.4.

16.8 Protein report

CLC Protein Workbench is able to produce protein reports, that allow you to easily generate different kinds of information regarding a protein.

Actually a protein report is a collection of some of the protein analyses which are described elsewhere in this manual.

To create a protein report do the following:

> **Right-click protein in Navigation Area** | **Toolbox** | **Protein Analyses** ( ) | **Create Protein Report** ( )

This opens dialog **Step 1**, where you can choose which proteins to create a report for. When the correct one is chosen, click **Next**.

In dialog **Step 2** you can choose which analyses you want to include in the report. The following list shows which analyses are available and explains where to find more details.

- **Sequence statistics.** See section 14.4 for more about this topic.
- **Plot of charge as function of pH.** See section 16.2 for more about this topic.
- **Plot of hydrophobicity.** See section 16.5 for more about this topic.
- **Plot of local complexity.** See section 14.3 for more about this topic.
- **Dot plot against self.** See section 14.2 for more about this topic.
- **Secondary structure prediction.** See section 16.7 for more about this topic.
- **Pfam domain search.** See section 16.6 for more about this topic.
- **Local BLAST.** See section 12.1.3 for more about this topic.
- **NCBI BLAST.** See section 12.1.1 for more about this topic.

When you have selected the relevant analyses, click **Next**. **Step 3** to **Step 7** (if you select all the analyses in **Step 2**) are adjustments of parameters for the different analyses. The parameters

are mentioned briefly in relation to the following steps, and you can turn to the relevant chapters or sections (mentioned above) to learn more about the significance of the parameters.

In **Step 3** you can adjust parameters for sequence statistics:

- **Individual Statistics Layout. Comparative** is disabled because reports are generated for one protein at a time.
- **Include Background Distribution of Amino Acids.** Includes distributions from different organisms. Background distributions are calculated from UniProt www.uniprot.org version 6.0, dated September 13 2005.

In **Step 4** you can adjust parameters for hydrophobicity plots:

- **Window size.** Width of window on sequence (odd number).
- **Hydrophobicity scales.** Lets you choose between different scales.

In **Step 5** you can adjust a parameter for complexity plots:

- **Window size.** Width of window on sequence (must be odd).

In **Step 6** you can adjust parameters for dot plots:

- **Score model.** Different scoring matrices.
- **Window size.** Width of window on sequence.

In **Step 7** you can adjust parameters for BLAST search:

- **Program.** Lets you choose between different BLAST programs.
- **Database.** Lets you limit your search to a particular database.

16.8.1 Protein report output

An example of Protein report can be seen in figure 16.20.

By double clicking a graph in the output, this graph is shown in a different view (*CLC Protein Workbench* generates another tab). The report output and the new graph views can be saved by dragging the tab into the **Navigation Area**.

The content of the tables in the report can be copy/pasted out of the program and e.g. into Microsoft Excel. To do so:

Select content of table | Right-click the selection | Copy

You can also **Export** ( ) the report in Excel format.

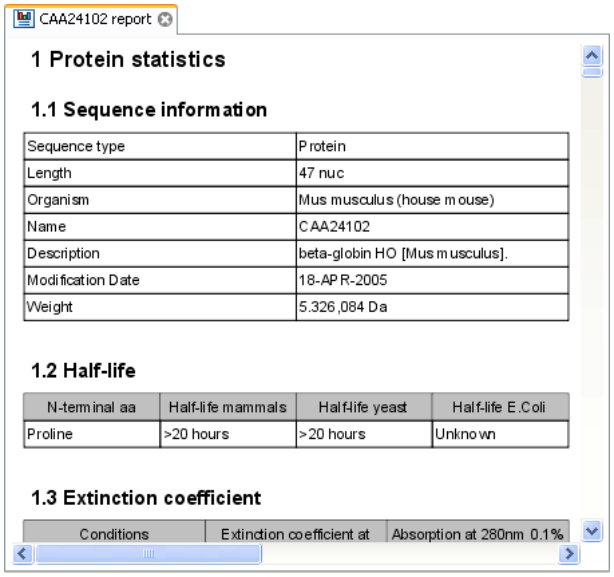

CAA24102 report
1 Protein statistics
1.1 Sequence information
Sequence type | Protein
Length | 47 nuc
Organism | Mus musculus (house mouse)
Name | CAA24102
Description | beta-globin HO [Mus musculus].
Modification Date | 18-APR-2005
Weight | 5.326,084 Da
1.2 Half-life
N-terminal aa | Half-life mammals | Half-life yeast | Half-life E.Coli
Proline | >20 hours | >20 hours | Unknown
1.3 Extinction coefficient
Conditions | Extinction coefficient at | Absorption at 280nm 0.1%

Figure 16.20: *A protein report. There is a Table of Contents in the Side Panel that makes it easy to browse the report.*

16.9 Reverse translation from protein into DNA

A protein sequence can be back-translated into DNA using *CLC Protein Workbench*. Due to degeneracy of the genetic code every amino acid could translate into several different codons (only 20 amino acids but 64 different codons). Thus, the program offers a number of choices for determining which codons should be used. These choices are explained in this section.

In order to make a reverse translation:

Select a protein sequence | **Toolbox in the Menu Bar** | **Protein Analyses ( )** | **Reverse Translate ( )**

or **right-click a protein sequence** | **Toolbox** | **Protein Analyses ( )** | **Reverse translate ()**

This opens the dialog displayed in figure 16.21:

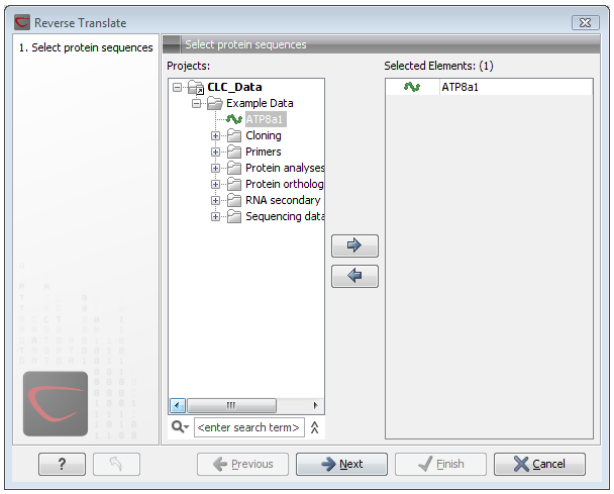


Figure 16.21: *Choosing a protein sequence for reverse translation.*

If a sequence was selected before choosing the Toolbox action, the sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements. You can translate several protein sequences at a time.

Click **Next** to adjust the parameters for the translation.

16.9.1 Reverse translation parameters

Figure 16.22 shows the choices for making the translation.

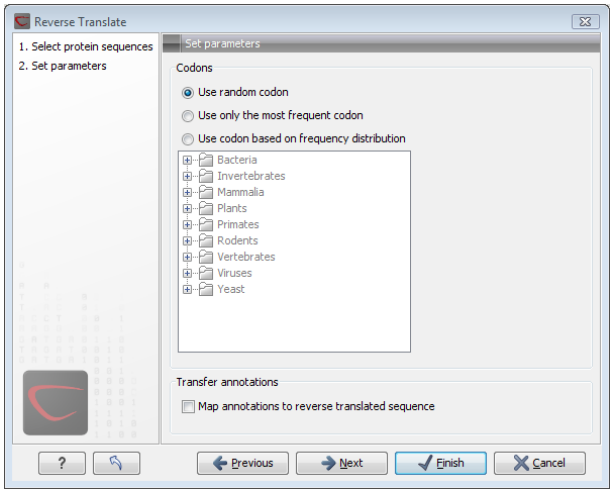


Figure 16.22: *Choosing parameters for the reverse translation.*

- **Use random codon.** This will randomly back-translate an amino acid to a codon without using the translation tables. Every time you perform the analysis you will get a different result.
- **Use only the most frequent codon.** On the basis of the selected translation table, this parameter/option will assign the codon that occurs most often. When choosing this option, the results of performing several reverse translations will always be the same, contrary to the other two options.
- **Use codon based on frequency distribution.** This option is a mix of the other two options. The selected translation table is used to attach weights to each codon based on its frequency. The codons are assigned randomly with a probability given by the weights. A more frequent codon has a higher probability of being selected. Every time you perform the analysis, you will get a different result. This option yields a result that is closer to the translation behavior of the organism (assuming you choose an appropriate codon frequency table).
- **Map annotations to reverse translated sequence.** If this checkbox is checked, then all annotations on the protein sequence will be mapped to the resulting DNA sequence. In the tooltip on the transferred annotations, there is a note saying that the annotation derives from the original sequence.

The **Codon Frequency Table** is used to determine the frequencies of the codons. Select a frequency table from the list that fits the organism you are working with. A translation table of an organism is created on the basis of counting all the codons in the coding sequences. Every codon in a **Codon Frequency Table** has its own count, frequency (per thousand) and fraction which are calculated in accordance with the occurrences of the codon in the organism. You can customize the list of codon frequency tables for your installation, see section I.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. The newly created nucleotide sequence is shown, and if the analysis was performed on several protein sequences, there will be a corresponding number of views of nucleotide sequences. The new sequence is not saved automatically. To save the sequence, drag it into the **Navigation Area** or press Ctrl + S (⌘ + S on Mac) to show the save dialog.

16.9.2 Bioinformatics explained: Reverse translation

In all living cells containing hereditary material such as DNA, a transcription to mRNA and subsequent a translation to proteins occur. This is of course simplified but is in general what is happening in order to have a steady production of proteins needed for the survival of the cell. In bioinformatics analysis of proteins it is sometimes useful to know the ancestral DNA sequence in order to find the genomic localization of the gene. Thus, the translation of proteins back to DNA/RNA is of particular interest, and is called reverse translation or back-translation.

The Genetic Code

In 1968 the Nobel Prize in Medicine was awarded to Robert W. Holley, Har Gobind Khorana and Marshall W. Nirenberg for their interpretation of the Genetic Code (http://nobelprize.org/medicine/laureates/1968/). The Genetic Code represents translations of all 64 different codons into 20 different amino acids. Therefore it is no problem to translate a DNA/RNA sequence into a specific protein. But due to the degeneracy of the genetic code, several codons may code for only one specific amino acid. This can be seen in the table below. After the discovery of the genetic code it has been concluded that different organism (and organelles) have genetic codes which are different from the "standard genetic code". Moreover, the amino acid alphabet is no longer limited to 20 amino acids. The 21'st amino acid, selenocysteine, is encoded by an 'UGA' codon which is normally a stop codon. The discrimination of a selenocysteine over a stop codon is carried out by the translation machinery. Selenocysteines are very rare amino acids.

The table below shows the Standard Genetic Code which is the default translation table.

TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys
TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys
TTA L Leu	TCA S Ser	TAA * Ter	TGA * Ter
TTG L Leu i	TCG S Ser	TAG * Ter	TGG W Trp
CTT L Leu	CCT P Pro	CAT H His	CGT R Arg
CTC L Leu	CCC P Pro	CAC H His	CGC R Arg
CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg
CTG L Leu i	CCG P Pro	CAG Q Gln	CGG R Arg
ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser
ATC I Ile	ACC T Thr	AAC N Asn	AGC S Ser
ATA I Ile	ACA T Thr	AAA K Lys	AGA R Arg
ATG M Met i	ACG T Thr	AAG K Lys	AGG R Arg
GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly
GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly
GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly
GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly

Challenge of reverse translation

A particular protein follows from the translation of a DNA sequence whereas the reverse translation need not have a specific solution according to the Genetic Code. The Genetic Code is degenerate which means that a particular amino acid can be translated into more than one codon. Hence there are ambiguities of the reverse translation.

Solving the ambiguities of reverse translation

In order to solve these ambiguities of reverse translation you can define how to prioritize the codon selection, e.g:

- Choose a codon randomly.
- Select the most frequent codon in a given organism.
- Randomize a codon, but with respect to its frequency in the organism.

As an example we want to translate an alanine to the corresponding codon. Four different codons can be used for this reverse translation; GCU, GCC, GCA or GCG. By picking either one by random choice we will get an alanine.

The most frequent codon, coding for an alanine in *E. coli* is GCG, encoding 33.7% of all alanines. Then comes GCC (25.5%), GCA (20.3%) and finally GCU (15.3%). The data are retrieved from the Codon usage database, see below. Always picking the most frequent codon does not necessarily give the best answer.

By selecting codons from a distribution of calculated codon frequencies, the DNA sequence obtained after the reverse translation, holds the correct (or nearly correct) codon distribution. It

should be kept in mind that the obtained DNA sequence is not necessarily identical to the original one encoding the protein in the first place, due to the degeneracy of the genetic code.

In order to obtain the best possible result of the reverse translation, one should use the codon frequency table from the correct organism or a closely related species. The codon usage of the mitochondrial chromosome are often different from the native chromosome(s), thus mitochondrial codon frequency tables should only be used when working specifically with mitochondria.

Other useful resources

The Genetic Code at NCBI:
http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c

Codon usage database:
http://www.kazusa.or.jp/codon/

Wikipedia on the genetic code
http://en.wikipedia.org/wiki/Genetic_code

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

16.10 Proteolytic cleavage detection

CLC Protein Workbench offers to analyze protein sequences with respect to cleavage by a selection of proteolytic enzymes. This section explains how to adjust the detection parameters and offers basic information on proteolytic cleavage in general.

16.10.1 Proteolytic cleavage parameters

Given a protein sequence, *CLC Protein Workbench* detects proteolytic cleavage sites in accordance with detection parameters and shows the detected sites as annotations on the sequence and in textual format in a table below the sequence view.

Detection of proteolytic cleavage sites is initiated by:

> **right-click a protein sequence in Navigation Area** | **Toolbox** | **Protein Analyses** ( ) | **Proteolytic Cleavage,** ( )

This opens the dialog shown in figure 16.23:

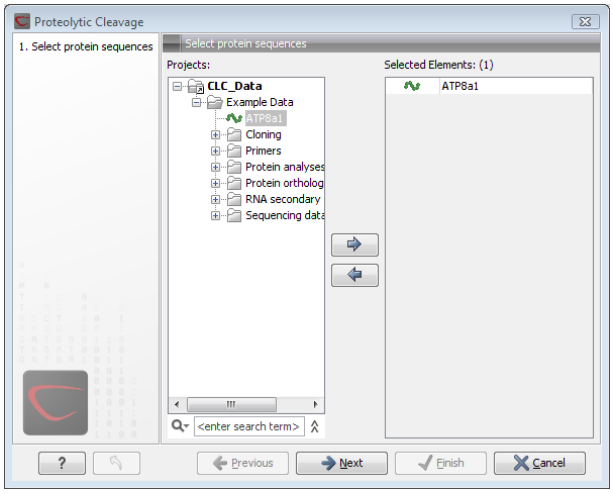


Figure 16.23: *Choosing sequence CAA32220 for proteolytic cleavage.*

CLC Protein Workbench allows you to detect proteolytic cleavages for several sequences at a time. Correct the list of sequences by selecting a sequence and clicking the arrows pointing left and right. Then click **Next** to go to **Step 2**.

In **Step 2** you can select proteolytic cleavage enzymes. The list of available enzymes will be expanded continuously. Presently, the list contains the enzymes shown in figure 16.24. The full list of enzymes and their cleavage patterns can be seen in Appendix, section E.

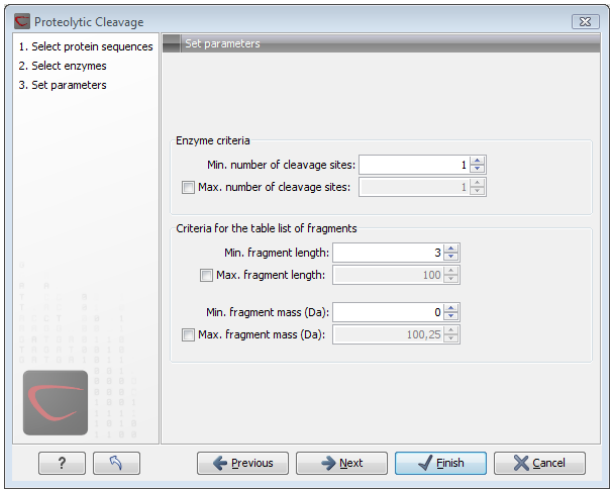


Figure 16.24: *Setting parameters for proteolytic cleavage detection.*

Select the enzymes you want to use for detection. When the relevant enzymes are chosen, click **Next**.

In **Step 3** you can set parameters for the detection. This limits the number of detected cleavages. Figure 16.25 shows an example of how parameters can be set.

- **Min. and max. number of cleavage sites.** Certain proteolytic enzymes cleave at many positions in the amino acid sequence. For instance proteinase K cleaves at nine different amino acids, regardless of the surrounding residues. Thus, it can be very useful to limit the number of actual cleavage sites before running the analysis.

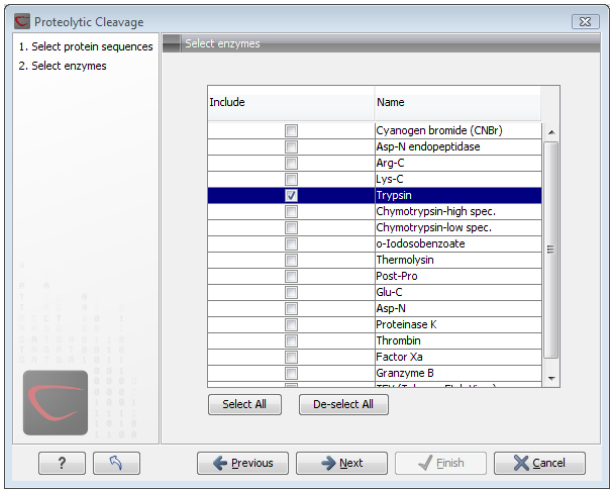


Figure 16.25: *Setting parameters for proteolytic cleavage detection.*

- **Min. and max. fragment length** Likewise, it is possible to limit the output to only display sequence fragments between a chosen length. Both a lower and upper limit can be chosen.
- **Min. and max. fragment mass** The molecular weight is not necessarily directly correlated to the fragment length as amino acids have different molecular masses. For that reason it is also possible to limit the search for proteolytic cleavage sites to mass-range.

Example!: If you have one protein sequence but you only want to show which enzymes cut between two and four times. Then you should select "The enzymes has more cleavage sites than 2" and select "The enzyme has less cleavage sites than 4". In the next step you should simply select all enzymes. This will result in a view where only enzymes which cut 2,3 or 4 times are presented.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

The result of the detection is displayed in figure 16.26.

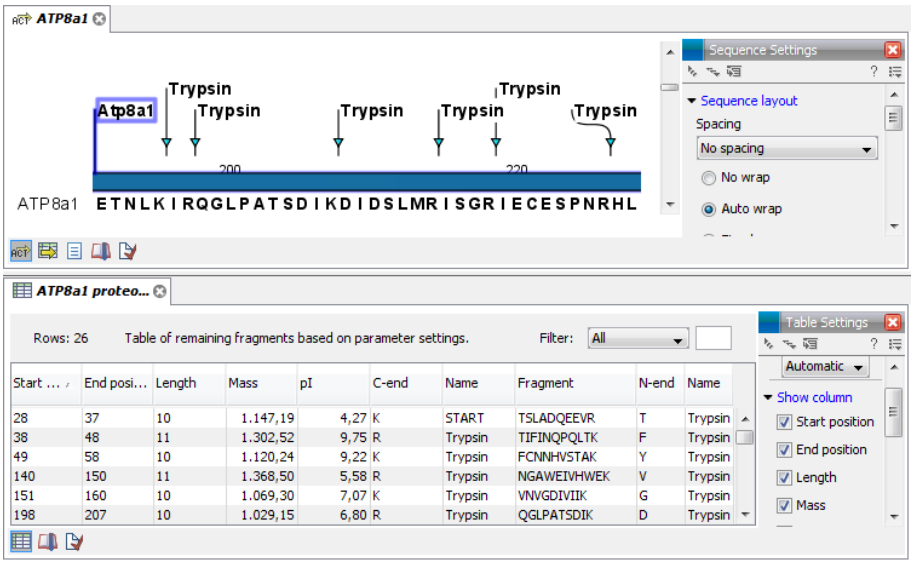

ATP8a1
Trypsin
Atp8a1
Trypsin
Trypsin
Trypsin
Trypsin
Trypsin
200
220
ATP8a1 ETNLKIRQGLPATSDIKDIDSLMRISGRIECESPNRHL
Sequence Settings
Sequence layout
Spacing
No spacing
No wrap
Auto wrap
ATP8a1 proteo...
Rows: 26
Table of remaining fragments based on parameter settings.
Filter: All
Table Settings
Automatic
Show column
Start position
End position
Length
Mass
Start ... | End posi... | Length | Mass | pI | C-end | Name | Fragment | N-end | Name
28 | 37 | 10 | 1.147,19 | 4,27 | K | START | TSLADQEEVR | T | Trypsin
38 | 48 | 11 | 1.302,52 | 9,75 | R | Trypsin | TIFINQPQLTK | F | Trypsin
49 | 58 | 10 | 1.120,24 | 9,22 | K | Trypsin | FCNNHVSTAK | Y | Trypsin
140 | 150 | 11 | 1.368,50 | 5,58 | R | Trypsin | NGAWEIVHWEK | V | Trypsin
151 | 160 | 10 | 1.069,30 | 7,07 | K | Trypsin | VNVGDIVIIK | G | Trypsin
198 | 207 | 10 | 1.029,15 | 6,80 | R | Trypsin | QGLPATSDIK | D | Trypsin

Figure 16.26: *The result of the proteolytic cleavage detection.*

Depending on the settings in the program, the output of the proteolytic cleavage site detection will display two views on the screen. The top view shows the actual protein sequence with the predicted cleavage sites indicated by small arrows. If no labels are found on the arrows they can be enabled by setting the labels in the "annotation layout" in the preference panel. The bottom view shows a text output of the detection, listing the individual fragments and information on these.

16.10.2 Bioinformatics explained: Proteolytic cleavage

Proteolytic cleavage is basically the process of breaking the peptide bonds between amino acids in proteins. This process is carried out by enzymes called peptidases, proteases or proteolytic cleavage enzymes.

Proteins often undergo proteolytic processing by specific proteolytic enzymes (proteases/peptidases) before final maturation of the protein. Proteins can also be cleaved as a result of intracellular processing of, for example, misfolded proteins. Another example of proteolytic processing of proteins is secretory proteins or proteins targeted to organelles, which have their signal peptide removed by specific signal peptidases before release to the extracellular environment or specific organelle.

Below a few processes are listed where proteolytic enzymes act on a protein substrate.

- N-terminal methionine residues are often removed after translation.
- Signal peptides or targeting sequences are removed during translocation through a membrane.
- Viral proteins that were translated from a monocistronic mRNA are cleaved.
- Proteins or peptides can be cleaved and used as nutrients.
- Precursor proteins are often processed to yield the mature protein.

Proteolytic cleavage of proteins has shown its importance in laboratory experiments where it is often useful to work with specific peptide fragments instead of entire proteins.

Proteases also have commercial applications. As an example proteases can be used as detergents for cleavage of proteinaceous stains in clothing.

The general nomenclature of cleavage site positions of the substrate were formulated by Schechter and Berger, 1967-68 [Schechter and Berger, 1967], [Schechter and Berger, 1968]. They designate the cleavage site between P1-P1', incrementing the numbering in the N-terminal direction of the cleaved peptide bond (P2, P3, P4, etc..). On the carboxyl side of the cleavage site the numbering is incremented in the same way (P1', P2', P3' etc.). This is visualized in figure 16.27.

Proteases often have a specific recognition site where the peptide bond is cleaved. As an example trypsin only cleaves at lysine or arginine residues, but it does not matter (with a few exceptions) which amino acid is located at position P1'(carboxyterminal of the cleavage site). Another example is trombin which cleaves if an arginine is found in position P1, but not if a D or E is found in position P1' at the same time. (See figure 16.28).

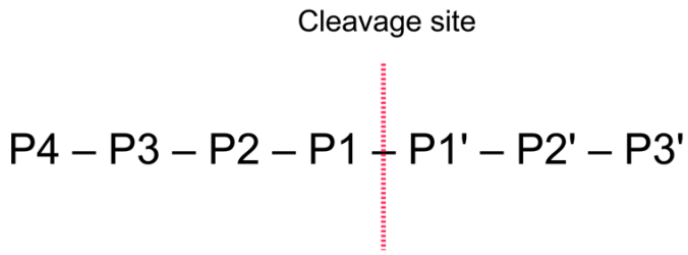


Figure 16.27: *Nomenclature of the peptide substrate. The substrate is cleaved between position P1-P1'.*

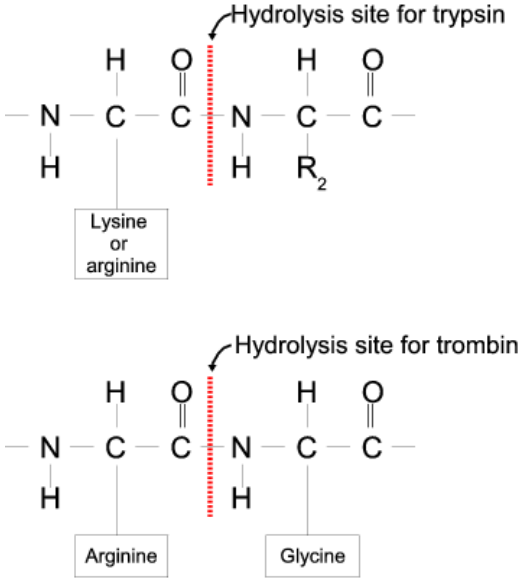


Figure 16.28: *Hydrolysis of the peptide bond between two amino acids. Trypsin cleaves unspecifically at lysine or arginine residues whereas trombin cleaves at arginines if asparate or glutamate is absent.*

Bioinformatics approaches are used to identify potential peptidase cleavage sites. Fragments can be found by scanning the amino acid sequence for patterns which match the corresponding cleavage site for the protease. When identifying cleaved fragments it is relatively important to know the calculated molecular weight and the isoelectric point.

Other useful resources

The Peptidase Database: http://merops.sanger.ac.uk/

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

Chapter 17

Restriction site analyses

Contents

17.1 Dynamic restriction sites . . . 266
17.1.1 Sort enzymes . . . 267
17.1.2 Manage enzymes . . . 268
17.2 Restriction site analysis from the Toolbox . . . 270
17.2.1 Selecting, sorting and filtering enzymes . . . 270
17.2.2 Number of cut sites . . . 271
17.2.3 Output of restriction map analysis . . . 272
17.2.4 Restriction sites as annotation on the sequence . . . 273
17.2.5 Table of restriction sites . . . 273
17.2.6 Table of restriction fragments . . . 274
17.2.7 Gel . . . 275
17.3 Gel electrophoresis . . . 275
17.3.1 Separate fragments of sequences on gel . . . 276
17.3.2 Separate sequences on gel . . . 276
17.3.3 Gel view . . . 276
17.4 Restriction enzyme lists . . . 278
17.4.1 Create enzyme list . . . 278
17.4.2 View and modify enzyme list . . . 280

There are two ways of finding and showing restriction sites:

- In many cases, the dynamic restriction sites found in the **Side Panel** of sequence views will be useful, since it is a quick and easy way of showing restriction sites.
- In the **Toolbox** you will find the other way of doing restriction site analyses. This way provides more control of the analysis and gives you more output options, e.g. a table of restriction sites and you can perform the same restriction map analysis on several sequences in one step.

This chapter first describes the dynamic restriction sites, followed by "the toolbox way". This section also includes an explanation of how to simulate a gel with the selected enzymes. The

final section in this chapter focuses on enzyme lists which represent an easy way of managing restriction enzymes.

17.1 Dynamic restriction sites

If you open a sequence, a sequence list etc, you will find the **Restriction Sites** group in the **Side Panel**.

As shown in figure 17.1 you can display restriction sites as colored triangles and lines on the sequence. The **Restriction sites** group in the side panel shows a list of enzymes, represented by different colors corresponding to the colors of the triangles on the sequence. By selecting or deselecting the enzymes in the list, you can specify which enzymes' restriction sites should be displayed.

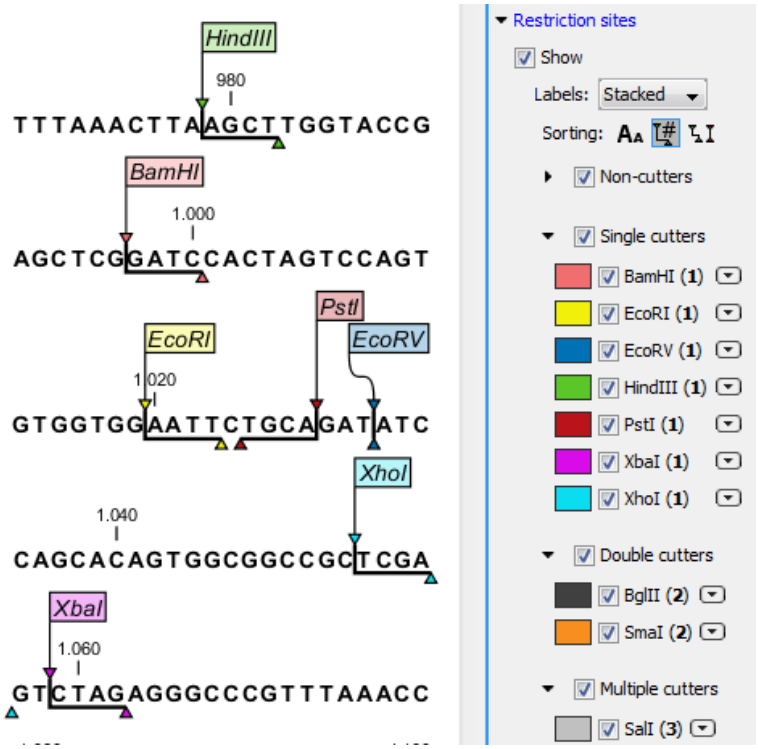


Figure 17.1: *Showing restriction sites of ten restriction enzymes.*

The color of the restriction enzyme can be changed by clicking the colored box next to the enzyme's name. The name of the enzyme can also be shown next to the restriction site by selecting **Show name flags** above the list of restriction enzymes.

There is also an option to specify how the **Labels** shown be shown:

- **No labels**. This will just display the cut site with no information about the name of the enzyme. Placing the mouse button on the cut site will reveal this information as a tool tip.
- **Flag**. This will place a flag just above the sequence with the enzyme name (see an example in figure 17.2). Note that this option will make it hard to see when several cut sites are located close to each other. In the circular view, this option is replaced by the Radial option:

- **Radial**. This option is only available in the circular view. It will place the restriction site labels as close to the cut site as possible (see an example in figure 17.4).
- **Stacked**. This is similar to the flag option for linear sequence views, but it will stack the labels so that all enzymes are shown. For circular views, it will align all the labels on each side of the circle. This can be useful for clearly seeing the order of the cut sites when they are located closely together (see an example in figure 17.3).

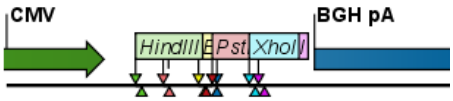


Figure 17.2: *Restriction site labels shown as flags.*

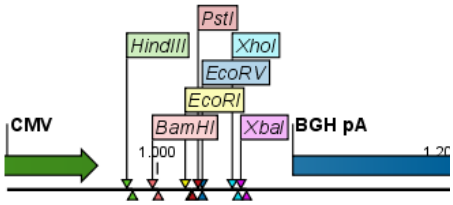


Figure 17.3: *Restriction site labels stacked.*

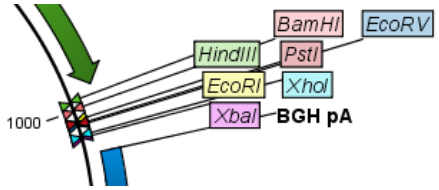


Figure 17.4: *Restriction site labels in radial layout.*

Note that in a circular view, the **Stacked** and **Radial** options also affect the layout of annotations.

17.1.1 Sort enzymes

Just above the list of enzymes there are three buttons to be used for sorting the list (see figure 17.5):

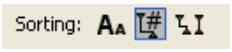


Figure 17.5: *Buttons to sort restriction enzymes.*

- **Sort enzymes alphabetically** (AA). Clicking this button will sort the list of enzymes alphabetically.
- **Sort enzymes by number of restriction sites** (#). This will divide the enzymes into four groups:
  - Non-cutters.
  - Single cutters.
  - Double cutters.
  - Multiple cutters.

There is a checkbox for each group which can be used to hide / show all the enzymes in a group.

-

- **Sort enzymes by overhang** ( ). This will divide the enzymes into three groups:
  - Blunt. Enzymes cutting both strands at the same position.
  - 3'. Enzymes producing an overhang at the 3' end.
  - 5'. Enzymes producing an overhang at the 5' end.

  There is a checkbox for each group which can be used to hide / show all the enzymes in a group.

17.1.2 Manage enzymes

The list of restriction enzymes contains per default 20 of the most popular enzymes, but you can easily modify this list and add more enzymes by clicking the **Manage enzymes button.** This will display the dialog shown in figure 17.6.

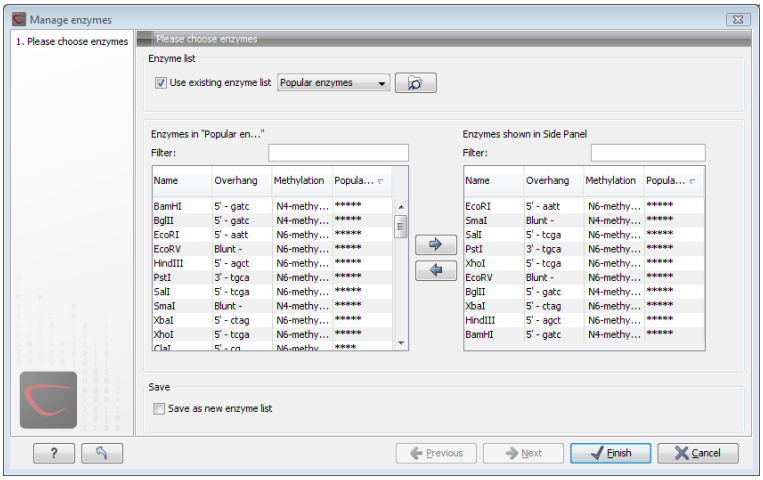


Figure 17.6: *Adding or removing enzymes from the Side Panel.*

At the top, you can choose to **Use existing enzyme list**. Clicking this option lets you select an enzyme list which is stored in the **Navigation Area**. See section 17.4 for more about creating and modifying enzyme lists.

Below there are two panels:

- To the **left**, you see all the enzymes that are in the list select above. If you have not chosen to use an existing enzyme list, this panel shows all the enzymes available [1].
- To the **right**, there is a list of the enzymes that will be used.

Select enzymes in the left side panel and add them to the right panel by double-clicking or clicking the **Add** button ( ). If you e.g. wish to use EcoRV and BamHI, select these two enzymes and add them to the right side panel.

[1] The *CLC Protein Workbench* comes with a standard set of enzymes based on http://www.rebase.neb.com. You can customize the enzyme database for your installation, see section **??**

If you wish to use all the enzymes in the list:

Click in the panel to the left | press Ctrl + A (⌘ + A on Mac) | Add (➡)

The enzymes can be sorted by clicking the column headings, i.e. Name, Overhang, Methylation or Popularity. This is particularly useful if you wish to use enzymes which produce e.g. a 3' overhang. In this case, you can sort the list by clicking the Overhang column heading, and all the enzymes producing 3' overhangs will be listed together for easy selection.

When looking for a specific enzyme, it is easier to use the Filter. If you wish to find e.g. HindIII sites, simply type HindIII into the filter, and the list of enzymes will shrink automatically to only include the HindIII enzyme. This can also be used to only show enzymes producing e.g. a 3' overhang as shown in figure 17.20.

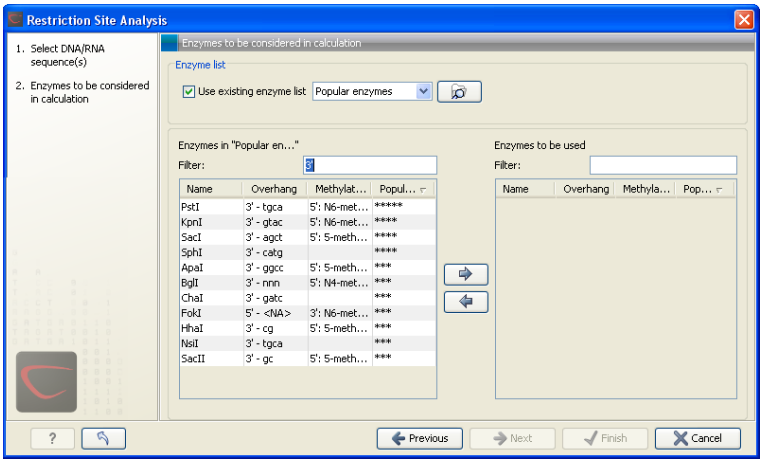

Restriction Site Analysis
1. Select DNA/RNA sequence(s)
2. Enzymes to be considered in calculation
Enzymes to be considered in calculation
Enzyme list
Use existing enzyme list
Popular enzymes
Enzymes in "Popular en..."
Filter:
Enzymes to be used
Name
Overhang
Methylat...
Popul...
PstI 3' - tgca 5': N6-met... *****
KpnI 3' - gtac 5': N6-met... ****
SacI 3' - agct 5': 5-meth... ****
SphI 3' - catg ****
ApaI 3' - ggcc 5': 5-meth... ***
BglI 3' - nnn 5': N4-met... ***
ClaI 3' - atc ***
FokI 5' - <NA> 3': N6-met... ***
HhaI 3' - cg 5': 5-meth... ***
NsiI 3' - tgca ***
SacII 3' - gc 5': 5-meth... ***
Previous
Next
Finish
Cancel

Figure 17.7: *Selecting enzymes.*

If you need more detailed information and filtering of the enzymes, either place your mouse cursor on an enzyme for one second to display additional information (see figure 17.21), or use the view of enzyme lists (see 17.4).

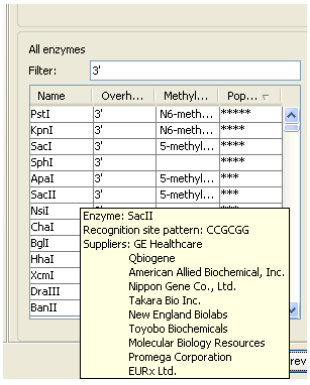


Figure 17.8: *Showing additional information about an enzyme like recognition sequence or a list of commercial vendors.*

At the bottom of the dialog, you can select to save this list of enzymes as a new file. In this way, you can save the selection of enzymes for later use.

When you click **Finish**, the enzymes are added to the Side Panel and the cut sites are shown on the sequence.

If you have specified a set of enzymes which you always use, it will probably be a good idea to save the settings in the Side Panel (see section 3.2.7) for future use.

17.2 Restriction site analysis from the Toolbox

Besides the dynamic restriction sites, you can do a more elaborate restriction map analysis with more output format using the Toolbox:

Toolbox | **Restriction Sites (** **)** | **Restriction Site Analysis (** **)**

This will display the dialog shown in figure 17.9.

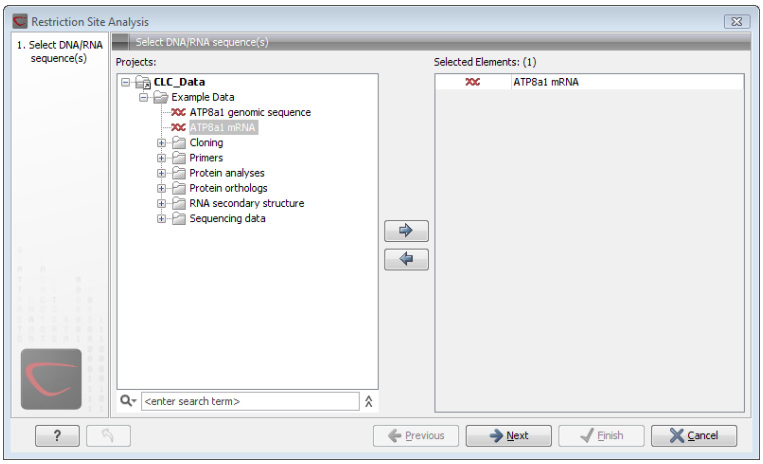


Figure 17.9: *Choosing sequence ATP8a1 mRNA for restriction map analysis.*

If a sequence was selected before choosing the Toolbox action, this sequence is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences or sequence lists from the selected elements.

17.2.1 Selecting, sorting and filtering enzymes

Clicking **Next** lets you define which enzymes to use as basis for finding restriction sites on the sequence. At the top, you can choose to **Use existing enzyme list**. Clicking this option lets you select an enzyme list which is stored in the **Navigation Area**. See section 17.4 for more about creating and modifying enzyme lists.

Below there are two panels:

- To the **left**, you see all the enzymes that are in the list select above. If you have not chosen to use an existing enzyme list, this panel shows all the enzymes available [2].
- To the **right**, there is a list of the enzymes that will be used.

Select enzymes in the left side panel and add them to the right panel by double-clicking or clicking the **Add** button ( ). If you e.g. wish to use EcoRV and BamHI, select these two enzymes and add them to the right side panel.

[2] The *CLC Protein Workbench* comes with a standard set of enzymes based on http://www.rebase.neb.com. You can customize the enzyme database for your installation, see section **??**

If you wish to use all the enzymes in the list:

Click in the panel to the left | press Ctrl + A (⌘ + A on Mac) | Add (➡)

The enzymes can be sorted by clicking the column headings, i.e. Name, Overhang, Methylation or Popularity. This is particularly useful if you wish to use enzymes which produce e.g. a 3' overhang. In this case, you can sort the list by clicking the Overhang column heading, and all the enzymes producing 3' overhangs will be listed together for easy selection.

When looking for a specific enzyme, it is easier to use the Filter. If you wish to find e.g. HindIII sites, simply type HindIII into the filter, and the list of enzymes will shrink automatically to only include the HindIII enzyme. This can also be used to only show enzymes producing e.g. a 3' overhang as shown in figure 17.20.

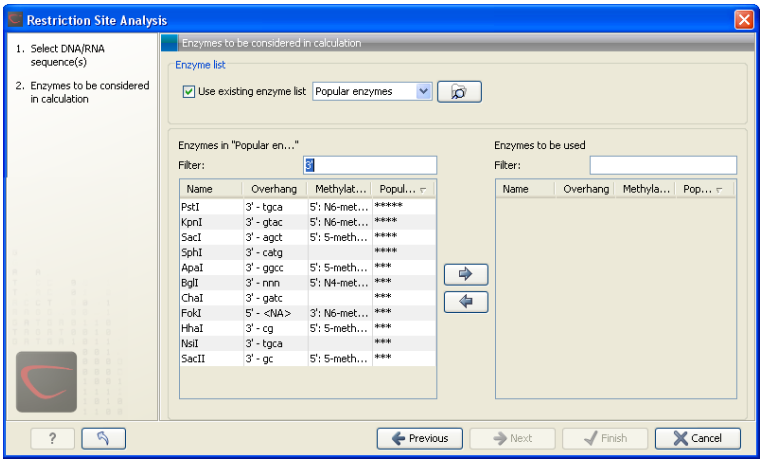

Restriction Site Analysis
1. Select DNA/RNA sequence(s)
2. Enzymes to be considered in calculation
Enzymes to be considered in calculation
Enzyme list
Use existing enzyme list
Popular enzymes
Enzymes in "Popular en..."
Filter:
Enzymes to be used
Name
Overhang
Methylat...
Popul...
PstI 3' - tgca 5': N6-met... *****
KpnI 3' - gtac 5': N6-met... ****
SacI 3' - agct 5': 5-meth... ****
SphI 3' - catg ****
ApaI 3' - ggcc 5': 5-meth... ***
BglI 3' - nnn 5': N4-met... ***
ClaI 3' - atc ***
FokI 5' - <NA> 3': N6-met... ***
HhaI 3' - cg 5': 5-meth... ***
NsiI 3' - tgca ***
SacII 3' - gc 5': 5-meth... ***
Previous
Next
Finish
Cancel

Figure 17.10: *Selecting enzymes.*

If you need more detailed information and filtering of the enzymes, either place your mouse cursor on an enzyme for one second to display additional information (see figure 17.21), or use the view of enzyme lists (see 17.4).

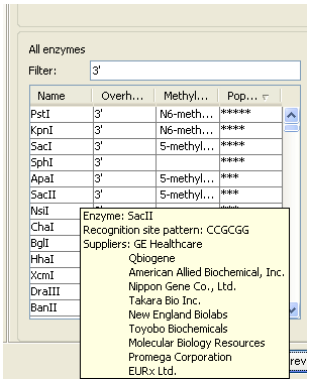


Figure 17.11: *Showing additional information about an enzyme like recognition sequence or a list of commercial vendors.*

17.2.2 Number of cut sites

Clicking **Next** confirms the list of enzymes which will be included in the analysis, and takes you to the dialog shown in figure 17.12.

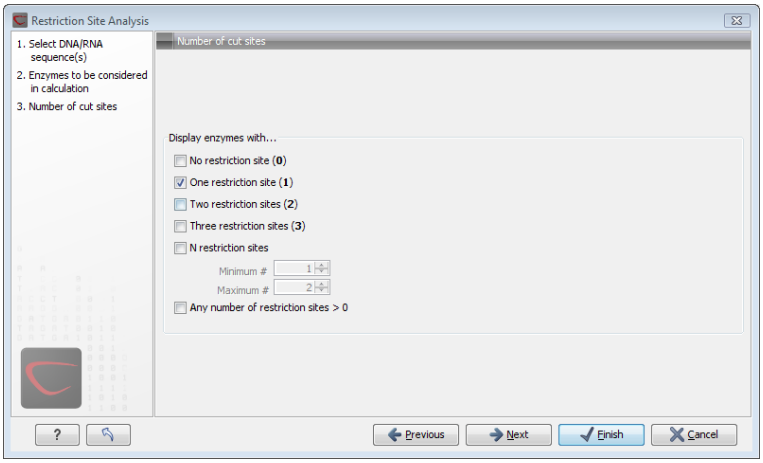


Figure 17.12: *Selecting number of cut sites.*

If you wish the output of the restriction map analysis only to include restriction enzymes which cut the sequence a specific number of times, use the checkboxes in this dialog:

- No restriction site (**0**)
- One restriction site (**1**)
- Two restriction sites (**2**)
- Three restriction site (**3**)
- N restriction sites
  - Minimum
  - Maximum
- Any number of restriction sites > 0

The default setting is to include the enzymes which cut the sequence one or two times.

You can use the checkboxes to perform very specific searches for restriction sites: e.g. if you wish to find enzymes which do not cut the sequence, or enzymes cutting exactly twice.

17.2.3 Output of restriction map analysis

Clicking next shows the dialog in figure 17.13.

This dialog lets you specify how the result of the restriction map analysis should be presented:

- **Add restriction sites as annotations to sequence(s)**. This option makes it possible to see the restriction sites on the sequence (see figure 17.14) and save the annotations for later use.
- **Create restriction map**. When a restriction map is created, it can be shown in three different ways:

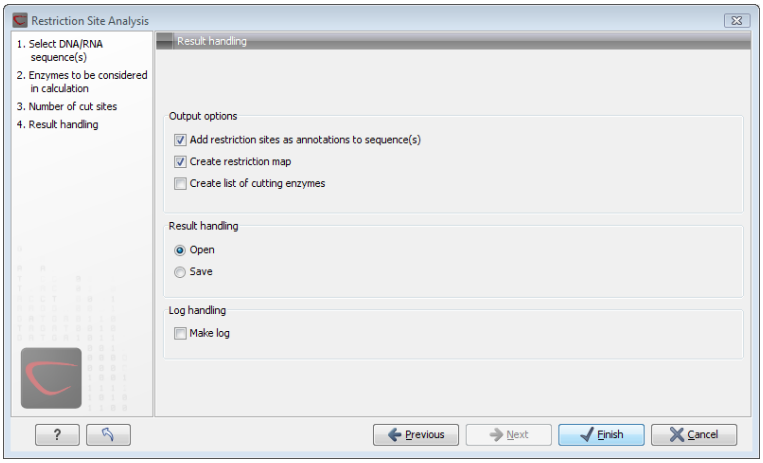


Figure 17.13: *Choosing to add restriction sites as annotations or creating a restriction map.*

- As a **table of restriction sites** as shown in figure 17.15. If more than one sequence were selected, the table will include the restriction sites of all the sequences. This makes it easy to compare the result of the restriction map analysis for two sequences.
- As a **table of fragments** which shows the sequence fragments that would be the result of cutting the sequence with the selected enzymes (see figure17.16).
- As a **virtual gel** simulation which shows the fragments as bands on a gel (see figure 17.18).

  For more information about gel electrophoresis, see section 17.3.

The following sections will describe these output formats in more detail.

In order to complete the analysis click **Finish** (see section 9.1 for information about the Save and Open options).

17.2.4 Restriction sites as annotation on the sequence

If you chose to add the restriction sites as annotation to the sequence, the result will be similar to the sequence shown in figure 17.14. See section 10.3 for more information about viewing

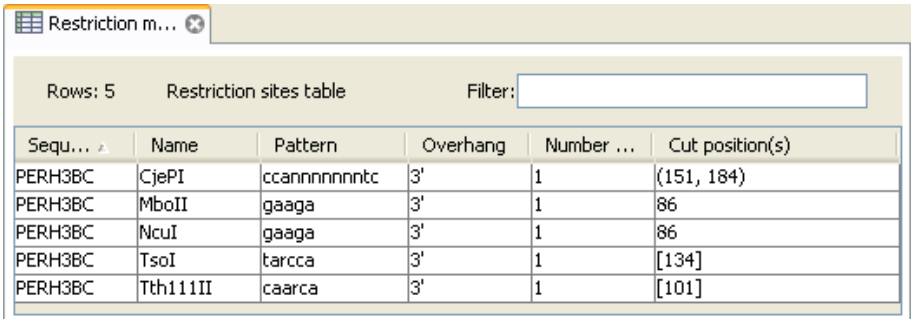
Figure 17.14: *The result of the restriction analysis shown as annotations.*

annotations.

17.2.5 Table of restriction sites

The restriction map can be shown as a table of restriction sites (see figure 17.15).

Each row in the table represents a restriction enzyme. The following information is available for each enzyme:



Sequ...	Name	Pattern	Overhang	Number ...	Cut position(s)
PERH3BC	CjePI	ccannnnnnntc	3'	1	(151, 184)
PERH3BC	MboII	gaaga	3'	1	86
PERH3BC	NcuI	gaaga	3'	1	86
PERH3BC	TsoI	tarcca	3'	1	[134]
PERH3BC	Tth111II	caarca	3'	1	[101]

Figure 17.15: *The result of the restriction analysis shown as annotations.*

- **Sequence**. The name of the sequence which is relevant if you have performed restriction map analysis on more than one sequence.
- **Name**. The name of the enzyme.
- **Pattern**. The recognition sequence of the enzyme.
- **Overhang**. The overhang produced by cutting with the enzyme (3', 5' or Blunt).
- **Number of cut sites**.
- **Cut position(s)**. The position of each cut.
  - **,** If the enzyme cuts more than once, the positions are separated by commas.
  - **[]** If the enzyme's recognition sequence is on the negative strand, the cut position is put in brackets (as the enzyme TsoI in figure 17.15 whose cut position is [134]).
  - **()** Some enzymes cut the sequence twice for each recognition site, and in this case the two cut positions are surrounded by parentheses.

17.2.6 Table of restriction fragments

The restriction map can be shown as a table of fragments produced by cutting the sequence with the enzymes:

Click the Fragments button ( ) at the bottom of the view

The table is shown in see figure 17.16.

Each row in the table represents a fragment. If more than one enzyme cuts in the same region, or if an enzyme's recognition site is cut by another enzyme, there will be a fragment for each of the possible cut combinations [3]. The following information is available for each fragment.

- **Sequence**. The name of the sequence which is relevant if you have performed restriction map analysis on more than one sequence.
- **Length**. The length of the fragment. If there are overhangs of the fragment, these are included in the length (both 3' and 5' overhangs).
- **Region**. The fragment's region on the original sequence.

[3] Furthermore, if this is the case, you will see the names of the other enzymes in the **Conflicting Enzymes** column

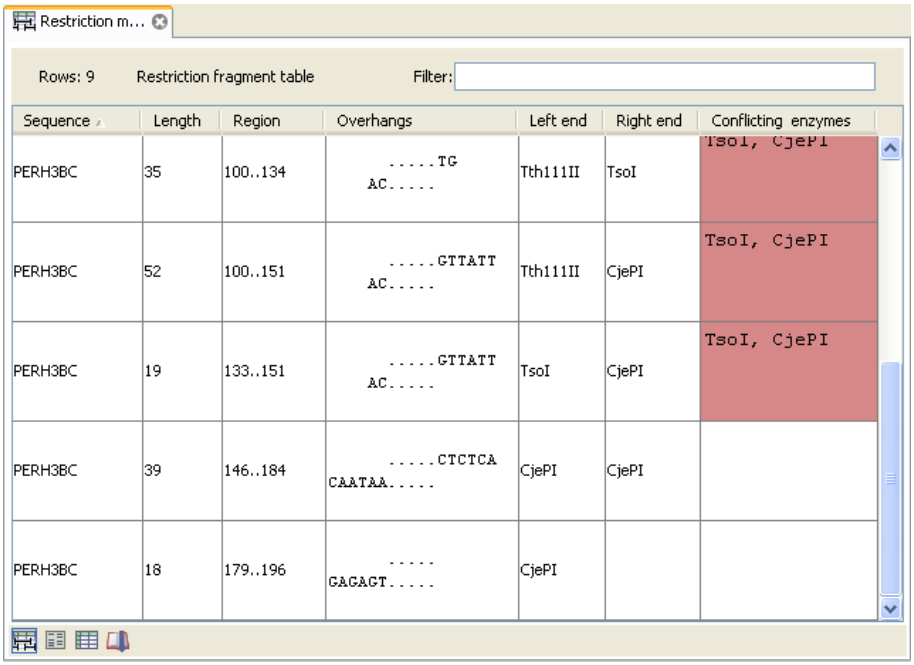


Figure 17.16: *The result of the restriction analysis shown as annotations.*

- **Overhangs**. If there is an overhang, this is displayed with an abbreviated version of the fragment and its overhangs. The two rows of dots (.) represent the two strands of the fragment and the overhang is visualized on each side of the dots with the residue(s) that make up the overhang. If there are only the two rows of dots, it means that there is no overhang.
- **Left end**. The enzyme that cuts the fragment to the left (5' end).
- **Right end**. The enzyme that cuts the fragment to the right (3' end).
- **Conflicting enzymes**. If more than one enzyme cuts at the same position, or if an enzyme's recognition site is cut by another enzyme, a fragment is displayed for each possible combination of cuts. At the same time, this column will display the enzymes that are in conflict. If there are conflicting enzymes, they will be colored red to alert the user. If the same experiment were performed in the lab, conflicting enzymes could lead to wrong results. For this reason, this functionality is useful to simulate digestions with complex combinations of restriction enzymes.

If views of both the fragment table and the sequence are open, clicking in the fragment table will select the corresponding region on the sequence.

17.2.7 Gel

The restriction map can also be shown as a gel. This is described in section 17.3.1.

17.3 Gel electrophoresis

CLC Protein Workbench enables the user to simulate the separation of nucleotide sequences on a gel. This feature is useful when e.g. designing an experiment which will allow the differentiation

of a successful and an unsuccessful cloning experiment on the basis of a restriction map.

There are two main ways to simulate gel separation of nucleotide sequences:

- One or more sequences can be digested with restriction enzymes and the resulting fragments can be separated on a gel.
- A number of existing sequences can be separated on a gel.

There are several ways to apply these functionalities as described below.

17.3.1 Separate fragments of sequences on gel

This section explains how to simulate a gel electrophoresis of one or more sequences which are digested with restriction enzymes. There are two ways to do this:

- When performing the **Restriction Site Analysis** from the **Toolbox**, you can choose to create a restriction map which can be shown as a gel. This is explained in section 17.2.
- From all the graphical views of sequences, you can right-click the name of the sequence and choose: **Digest Sequence with Selected Enzymes and Run on Gel ( )**. The views where this option is available are listed below:
  - Circular view (see section 10.2).
  - Ordinary sequence view (see section 10.1).
  - Graphical view of sequence lists (see section 10.7).

  Furthermore, you can also right-click an empty part of the view of the graphical view of sequence lists and choose **Digest All Sequences with Selected Enzymes and Run on Gel**. **Note!** When using the right-click options, the sequence will be digested with the enzymes that are selected in the **Side Panel**. This is explained in section 10.1.2.

The view of the gel is explained in section 17.3.3

17.3.2 Separate sequences on gel

To separate sequences without restriction enzyme digestion, first create a sequence list of the sequences in question (see section 10.7). Then click the **Gel** button ( ) at the bottom of the view of the sequence list.

For more information about the view of the gel, see the next section.

17.3.3 Gel view

In figure 17.18 you can see a simulation of a gel with its **Side Panel** to the right. This view will be explained in this section.

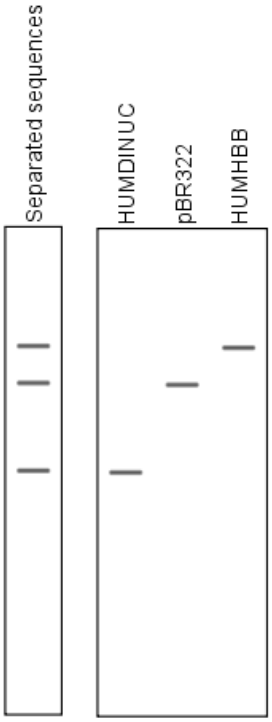


Figure 17.17: *A sequence list shown as a gel.*

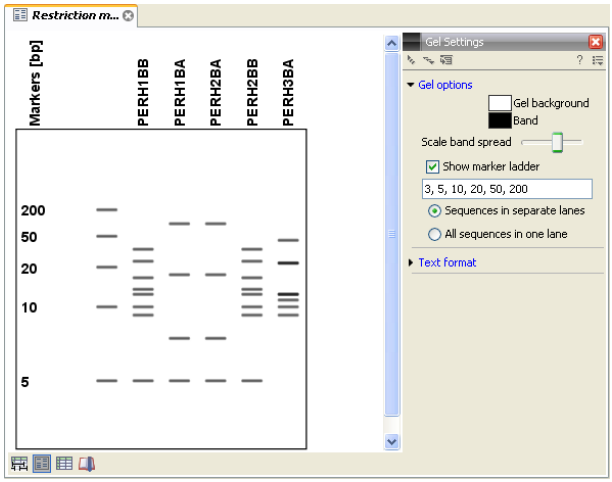


Figure 17.18: *Five lanes showing fragments of five sequences cut with restriction enzymes.*

Information on bands / fragments

You can get information about the individual bands by hovering the mouse cursor on the band of interest. This will display a tool tip with the following information:

- Fragment length
- Fragment region on the original sequence
- Enzymes cutting at the left and right ends, respectively

For gels comparing whole sequences, you will see the sequence name and the length of the sequence.

Note! You have to be in **Selection** ( ) or **Pan** ( ) mode in order to get this information.

It can be useful to add markers to the gel which enables you to compare the sizes of the bands. This is done by clicking **Show marker ladder** in the **Side Panel**.

Markers can be entered into the text field, separated by commas.

Modifying the layout

The background of the lane and the colors of the bands can be changed in the **Side Panel**. Click the colored box to display a dialog for picking a color. The slider **Scale band spread** can be used to adjust the effective time of separation on the gel, i.e. how much the bands will be spread over the lane. In a real electrophoresis experiment this property will be determined by several factors including time of separation, voltage and gel density.

You can also choose how many lanes should be displayed:

- **Sequences in separate lanes**. This simulates that a gel is run for each sequence.
- **All sequences in one lane**. This simulates that one gel is run for all sequences.

You can also modify the layout of the view by zooming in or out. Click **Zoom in** ( ) or **Zoom out** ( ) in the Toolbar and click the view.

Finally, you can modify the format of the text heading each lane in the **Text format** preferences in the **Side Panel**.

17.4 Restriction enzyme lists

CLC Protein Workbench includes all the restriction enzymes available in the **REBASE** database[4]. However, when performing restriction site analyses, it is often an advantage to use a customized list of enzymes. In this case, the user can create special lists containing e.g. all enzymes available in the laboratory freezer, all enzymes used to create a given restriction map or all enzymes that are available form the preferred vendor.

In the example data (see section 1.6.2) under Nucleotide->Restriction analysis, there are two enzyme lists: one with the 50 most popular enzymes, and another with all enzymes that are included in the *CLC Protein Workbench*.

This section describes how you can create an enzyme list, and how you can modify it.

17.4.1 Create enzyme list

CLC Protein Workbench uses enzymes from the **REBASE** restriction enzyme database at `http://rebase.neb.com`[5].

To create an enzyme list of a subset of these enzymes:

[4] You can customize the enzyme database for your installation, see section **??**
[5] You can customize the enzyme database for your installation, see section **??**

File | New | Enzyme list ( )

This opens the dialog shown in figure 17.19

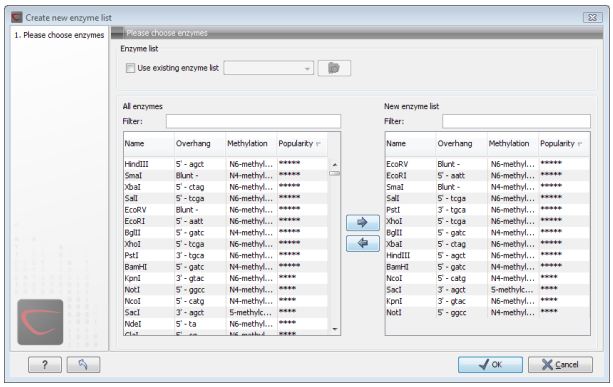


Figure 17.19: *Choosing enzymes for the new enzyme list.*

At the top, you can choose to **Use existing enzyme list**. Clicking this option lets you select an enzyme list which is stored in the **Navigation Area**. See section 17.4 for more about creating and modifying enzyme lists.

Below there are two panels:

- To the **left**, you see all the enzymes that are in the list select above. If you have not chosen to use an existing enzyme list, this panel shows all the enzymes available [6].
- To the **right**, there is a list of the enzymes that will be used.

Select enzymes in the left side panel and add them to the right panel by double-clicking or clicking the **Add** button ( ). If you e.g. wish to use EcoRV and BamHI, select these two enzymes and add them to the right side panel.

If you wish to use all the enzymes in the list:

Click in the panel to the left | press Ctrl + A (⌘ + A on Mac) | Add ( )

The enzymes can be sorted by clicking the column headings, i.e. Name, Overhang, Methylation or Popularity. This is particularly useful if you wish to use enzymes which produce e.g. a 3' overhang. In this case, you can sort the list by clicking the Overhang column heading, and all the enzymes producing 3' overhangs will be listed together for easy selection.

When looking for a specific enzyme, it is easier to use the Filter. If you wish to find e.g. HindIII sites, simply type HindIII into the filter, and the list of enzymes will shrink automatically to only include the HindIII enzyme. This can also be used to only show enzymes producing e.g. a 3' overhang as shown in figure 17.20.

If you need more detailed information and filtering of the enzymes, either place your mouse cursor on an enzyme for one second to display additional information (see figure 17.21), or use the view of enzyme lists (see 17.4).

Click **Finish** to open the enzyme list.

[6] The *CLC Protein Workbench* comes with a standard set of enzymes based on http://www.rebase.neb.com. You can customize the enzyme database for your installation, see section **??**

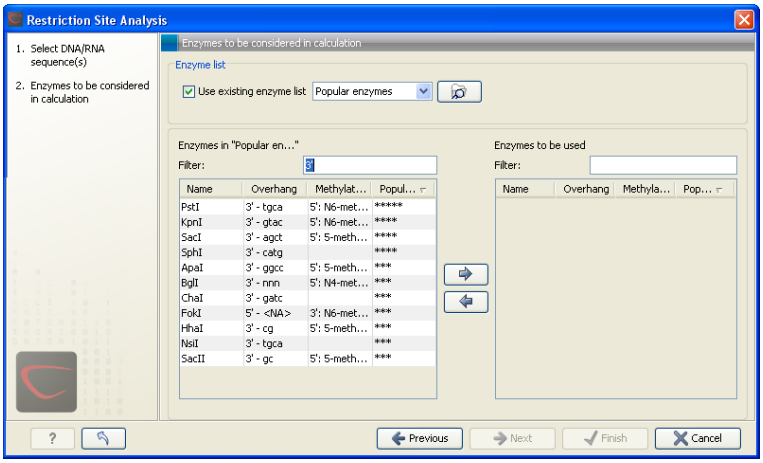


Figure 17.20: *Selecting enzymes.*

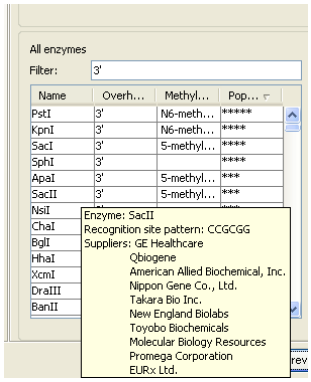


Figure 17.21: *Showing additional information about an enzyme like recognition sequence or a list of commercial vendors.*

17.4.2 View and modify enzyme list

An enzyme list is shown in figure 17.22. The list can be sorted by clicking the columns,

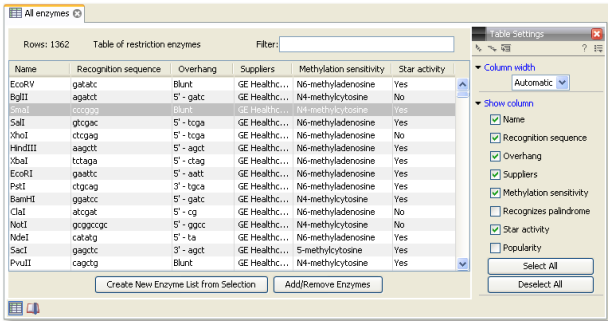

All enzymes
Rows: 1362
Table of restriction enzymes
Filter:
Name
Recognition sequence
Overhang
Suppliers
Methylation sensitivity
Star activity
EcoRV gatatc Blunt GE Healthc... N6-methyladenosine Yes
BglII agatct 5' - gatc GE Healthc... N4-methylcytosine No
ClaI atcgat 5' - cg GE Healthc... N6-methyladenosine No
SalI gtcgac 5' - tcga GE Healthc... N6-methyladenosine Yes
XhoI ctcgag 5' - tcga GE Healthc... N6-methyladenosine No
HindIII aagctt 5' - agct GE Healthc... N6-methyladenosine Yes
XbaI tctaga 5' - ctag GE Healthc... N6-methyladenosine Yes
EcoRI gaattc 5' - aatt GE Healthc... N6-methyladenosine Yes
PstI ctgcag 3' - tgca GE Healthc... N6-methyladenosine Yes
BamHI ggatcc 5' - gatc GE Healthc... N4-methylcytosine Yes
NotI gcggccgc 5' - ggcc GE Healthc... N4-methylcytosine No
NdeI catatg 5' - ta GE Healthc... N6-methyladenosine Yes
SacI gagctc 3' - agct GE Healthc... 5-methylcytosine Yes
PvuII cagctg Blunt GE Healthc... N4-methylcytosine Yes
Create New Enzyme List from Selection
Add/Remove Enzymes
Table Settings
Column width
Automatic
Show column
Name
Recognition sequence
Overhang
Suppliers
Methylation sensitivity
Recognizes palindrome
Star activity
Popularity
Select All
Deselect All

Figure 17.22: *An enzyme list.*

and you can use the filter at the top right corner to search for specific enzymes, recognition sequences etc.

If you wish to remove or add enzymes, click the **Add/Remove Enzymes** button at the bottom of the view. This will present the same dialog as shown in figure 17.19 with the enzyme list shown to the right.

If you wish to extract a subset of an enzyme list:

open the list | select the relevant enzymes | right-click | Create New Enzyme List from Selection ( )

If you combined this method with the filter located at the top of the view, you can extract a very specific set of enzymes. E.g. if you wish to create a list of enzymes sold by a particular distributor, type the name of the distributor into the filter, and select and create a new enzyme list from the selection.

Chapter 18

Sequence alignment

Contents

18.1 Create an alignment . . . 283
18.1.1 Gap costs . . . 284
18.1.2 Fast or accurate alignment algorithm . . . 284
18.1.3 Aligning alignments . . . 285
18.1.4 Fixpoints . . . 286
18.2 View alignments . . . 288
18.2.1 Bioinformatics explained: Sequence logo . . . 290
18.3 Edit alignments . . . 292
18.3.1 Move residues and gaps . . . 292
18.3.2 Insert gaps . . . 292
18.3.3 Delete residues and gaps . . . 292
18.3.4 Copy annotations to other sequences . . . 293
18.3.5 Move sequences up and down . . . 293
18.3.6 Delete, rename and add sequences . . . 293
18.3.7 Realign selection . . . 294
18.4 Join alignments . . . 294
18.4.1 How alignments are joined . . . 296
18.5 Pairwise comparison . . . 296
18.5.1 Pairwise comparison on alignment selection . . . 296
18.5.2 Pairwise comparison parameters . . . 297
18.5.3 The pairwise comparison table . . . 298
18.6 Bioinformatics explained: Multiple alignments . . . 299
18.6.1 Use of multiple alignments . . . 299
18.6.2 Constructing multiple alignments . . . 299

CLC Protein Workbench can align nucleotides and proteins using a *progressive alignment* algorithm (see section 18.6 or read the White paper on alignments in the **Science** section of http://www.clcbio.com).

This chapter describes how to use the program to align sequences. The chapter also describes alignment algorithms in more general terms.

18.1 Create an alignment

Alignments can be created from sequences, sequence lists (see section 10.7), existing alignments and from any combination of the three.

To create an alignment in *CLC Protein Workbench*:

select sequences to align | **Toolbox in the Menu Bar** | **Alignments and Trees** ( ) | **Create Alignment** ( )

or **select sequences to align** | **right-click any selected sequence** | **Toolbox** | **Alignments and Trees** ( ) | **Create Alignment** ( )

This opens the dialog shown in figure 18.1.

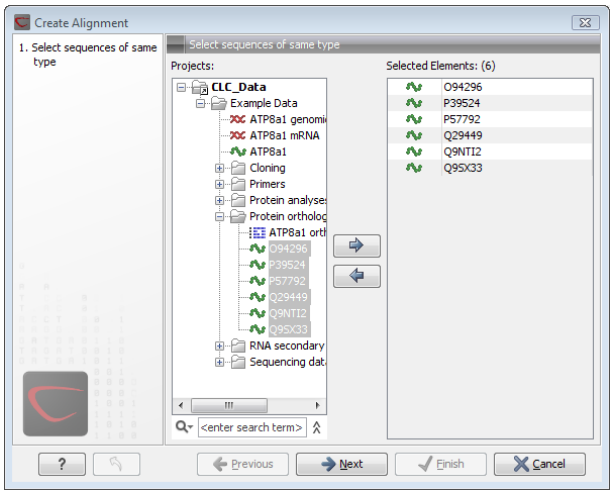


Figure 18.1: *Creating an alignment.*

If you have selected some elements before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove sequences, sequence lists or alignments from the selected elements. Click **Next** to adjust alignment algorithm parameters. Clicking **Next** opens the dialog shown in figure 18.2.

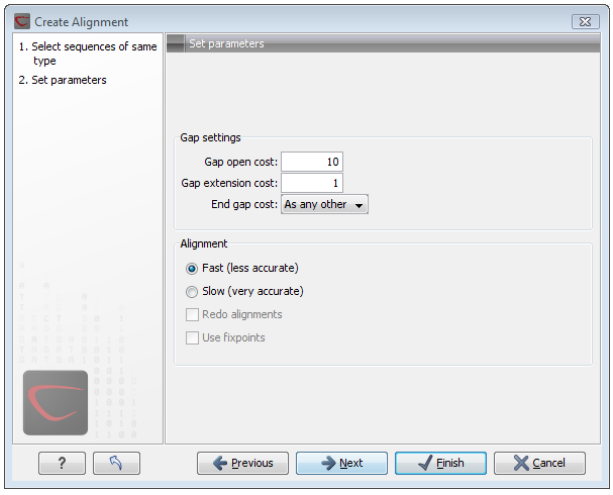


Figure 18.2: *Adjusting alignment algorithm parameters.*

18.1.1 Gap costs

The alignment algorithm has three parameters concerning gap costs: Gap open cost, Gap extension cost and End gap cost. The precision of these parameters is to one place of decimal.

- **Gap open cost**. The price for introducing gaps in an alignment.
- **Gap extension cost**. The price for every extension past the initial gap.

If you expect a lot of small gaps in your alignment, the Gap open cost should equal the Gap extension cost. On the other hand, if you expect few but large gaps, the Gap open cost should be set significantly higher than the Gap extension cost.

However, for most alignments it is a good idea to make the Gap open cost quite a bit higher than the Gap extension cost. The default values are 10.0 and 1.0 for the two parameters, respectively.

- **End gap cost**. The price of gaps at the beginning or the end of the alignment. One of the advantages of the *CLC Protein Workbench* alignment method is that it provides flexibility in the treatment of gaps at the ends of the sequences. There are three possibilities:
  - **Free end gaps**. Any number of gaps can be inserted in the ends of the sequences without any cost.
  - **Cheap end gaps**. All end gaps are treated as gap extensions and any gaps past 10 are free.
  - **End gaps as any other**. Gaps at the ends of sequences are treated like gaps in any other place in the sequences.

When aligning a long sequence with a short partial sequence, it is ideal to use free end gaps, since this will be the best approximation to the situation. The many gaps inserted at the ends are not due to evolutionary events, but rather to partial data.

Many homologous proteins have quite different ends, often with large insertions or deletions. This confuses alignment algorithms, but using the **Cheap end gaps** option, large gaps will generally be tolerated at the sequence ends, improving the overall alignment. This is the default setting of the algorithm.

Finally, treating end gaps like any other gaps is the best option when you know that there are no biologically distinct effects at the ends of the sequences.

Figures 18.3 and 18.4 illustrate the differences between the different gap scores at the sequence ends.

18.1.2 Fast or accurate alignment algorithm

CLC Protein Workbench has two algorithms for calculating alignments:

- **Fast (less accurate).** This allows for use of an optimized alignment algorithm which is very fast. The fast option is particularly useful for data sets with very long sequences.
- **Slow (very accurate).** This is the recommended choice unless you find the processing time too long.

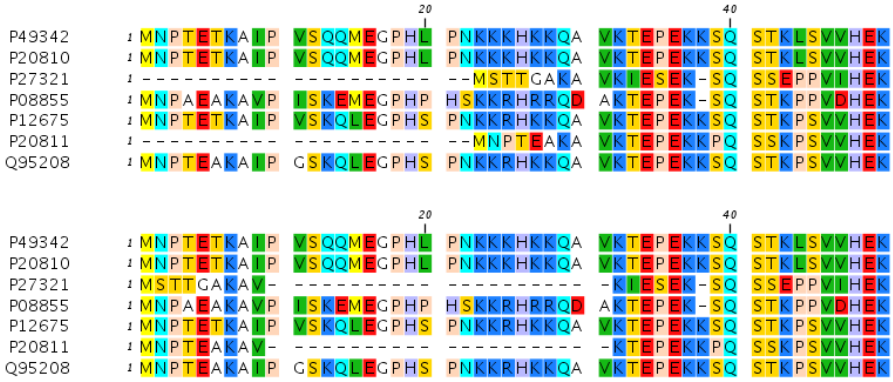

20
40
P49342 1 MNPTETKAIP VSQQMEGPHL PNKKKHKKQA VKTEPEKKSQ STKLSVVHEK
P20810 1 MNPTETKAIP VSQQMEGPHL PNKKKHKKQA VKTEPEKKSQ STKLSVVHEK
P27321 1 ---------- ---------- --MSTTGAKA VKIESEK-SQ SSEPPVIHEK
P08855 1 MNPAEAKAVP ISKEMEGPHP HSKKRHRRQD AKTEPEK-SQ STKPPVDHEK
P12675 1 MNPTETKAIP VSKQLEGPHS PNKKRHKKQA VKTEPEKKSQ STKPSVVHEK
P20811 1 ---------- ---------- --MNPTEAKA VKTEPEKKPQ SSKPSVVHEK
Q95208 1 MNPTEAKAIP GSKQLEGPHS PNKKRHKKQA VKTEPEKKSQ STKPSVVHEK
20
40
P49342 1 MNPTETKAIP VSQQMEGPHL PNKKKHKKQA VKTEPEKKSQ STKLSVVHEK
P20810 1 MNPTETKAIP VSQQMEGPHL PNKKKHKKQA VKTEPEKKSQ STKLSVVHEK
P27321 1 MSTTGAKAV- ---------- ---------- -KIESEK-SQ SSEPPVIHEK
P08855 1 MNPAEAKAVP ISKEMEGPHP HSKKRHRRQD AKTEPEK-SQ STKPPVDHEK
P12675 1 MNPTETKAIP VSKQLEGPHS PNKKRHKKQA VKTEPEKKSQ STKPSVVHEK
P20811 1 MNPTEAKAV- ---------- ---------- -KTEPEKKPQ SSKPSVVHEK
Q95208 1 MNPTEAKAIP GSKQLEGPHS PNKKRHKKQA VKTEPEKKSQ STKPSVVHEK

Figure 18.3: *The first 50 positions of two different alignments of seven calpastatin sequences. The top alignment is made with cheap end gaps, while the bottom alignment is made with end gaps having the same price as any other gaps. In this case it seems that the latter scoring scheme gives the best result.*

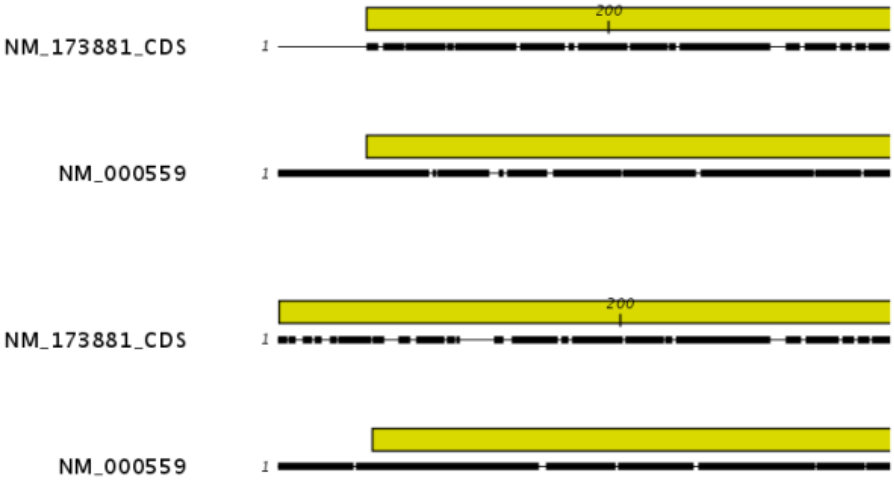


Figure 18.4: *The alignment of the coding sequence of bovine myoglobin with the full mRNA of human gamma globin. The top alignment is made with free end gaps, while the bottom alignment is made with end gaps treated as any other. The yellow annotation is the coding sequence in both sequences. It is evident that free end gaps are ideal in this situation as the start codons are aligned correctly in the top alignment. Treating end gaps as any other gaps in the case of aligning distant homologs where one sequence is partial leads to a spreading out of the short sequence as in the bottom alignment.*

Both algorithms use progressive alignment. The faster algorithm builds the initial tree by doing more approximate pairwise alignments than the slower option.

18.1.3 Aligning alignments

If you have selected an existing alignment in the first step (18.1), you have to decide how this alignment should be treated.

- **Redo alignment.** The original alignment will be realigned if this checkbox is checked. Otherwise, the original alignment is kept in its original form except for possible extra equally sized gaps in all sequences of the original alignment. This is visualized in figure 18.5.

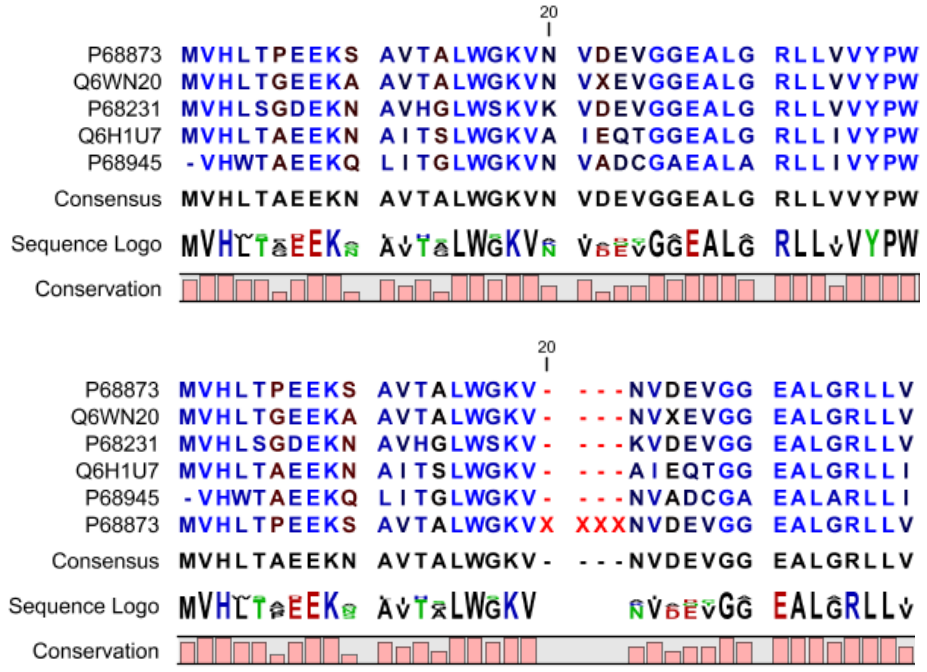


Figure 18.5: *The top figures shows the original alignment. In the bottom panel a single sequence with four inserted X's are aligned to the original alignment. This introduces gaps in all sequences of the original alignment. All other positions in the original alignment are fixed.*

This feature is useful if you wish to add extra sequences to an existing alignment, in which case you just select the alignment and the extra sequences and choose not to redo the alignment.

It is also useful if you have created an alignment where the gaps are not placed correctly. In this case, you can realign the alignment with different gap cost parameters.

18.1.4 Fixpoints

With fixpoints, you can get full control over the alignment algorithm. The fixpoints are points on the sequences that are forced to align to each other.

Fixpoints are added to sequences or alignments before clicking "Create alignment". To add a fixpoint, open the sequence or alignment and:

> **Select the region you want to use as a fixpoint | right-click the selection | Set alignment fixpoint here**

This will add an annotation labeled "Fixpoint" to the sequence (see figure 18.6). Use this procedure to add fixpoints to the other sequence(s) that should be forced to align to each other.

When you click "Create alignment" and go to **Step 2**, check **Use fixpoints** in order to force the alignment algorithm to align the fixpoints in the selected sequences to each other.

In figure 18.7 the result of an alignment using fixpoints is illustrated.

You can add multiple fixpoints, e.g. adding two fixpoints to the sequences that are aligned will force their first fixpoints to be aligned to each other, and their second fixpoints will also be

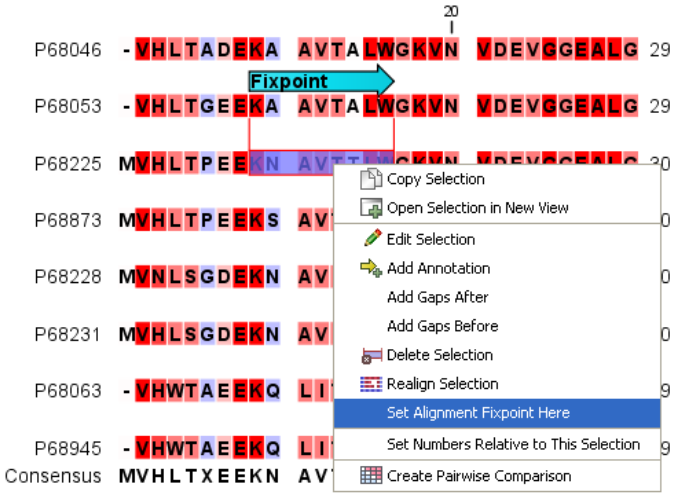


Figure 18.6: *Adding a fixpoint to a sequence in an existing alignment. At the top you can see a fixpoint that has already been added.*

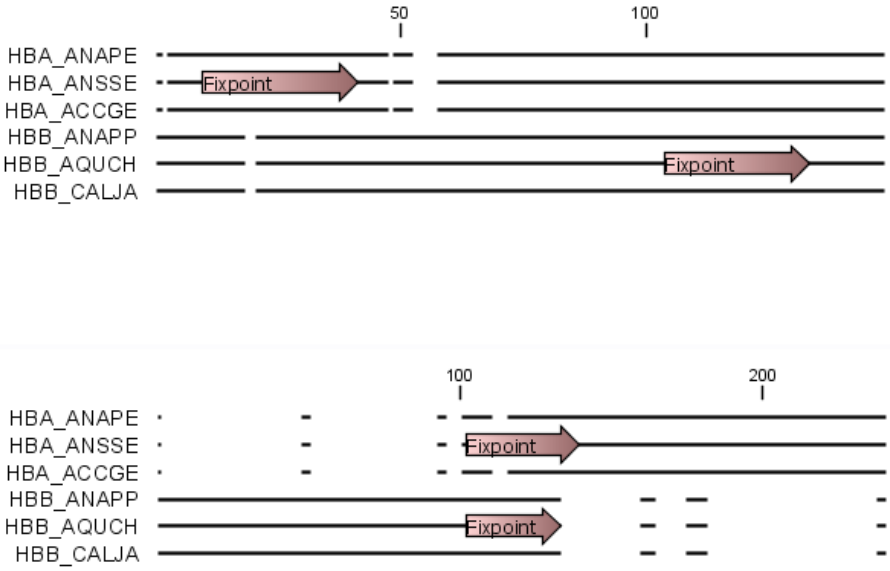


Figure 18.7: *Realigning using fixpoints. In the top view, fixpoints have been added to two of the sequences. In the view below, the alignment has been realigned using the fixpoints. The three top sequences are very similar, and therefore they follow the one sequence (number two from the top) that has a fixpoint.*

aligned to each other.

Advanced use of fixpoints

Fixpoints with the same names will be aligned to each other, which gives the opportunity for great control over the alignment process. It is only necessary to change any fixpoint names in very special cases.

One example would be three sequences A, B and C where sequences A and B has one copy of a domain while sequence C has two copies of the domain. You can now force sequence A to align to the first copy and sequence B to align to the second copy of the domains in sequence C. This is done by inserting fixpoints in sequence C for each domain, and naming them ’fp1’ and ’fp2’

(for example). Now, you can insert a fixpoint in each of sequences A and B, naming them 'fp1' and 'fp2', respectively. Now, when aligning the three sequences using fixpoints, sequence A will align to the first copy of the domain in sequence C, while sequence B would align to the second copy of the domain in sequence C.

You can name fixpoints by:

> **right-click the Fixpoint annotation | Edit Annotation ( ) | type the name in the 'Name' field**

18.2 View alignments

Since an alignment is a display of several sequences arranged in rows, the basic options for viewing alignments are the same as for viewing sequences. Therefore we refer to section 10.1 for an explanation of these basic options.

However, there are a number of alignment-specific view options in the **Alignment info** and the **Nucleotide info** in the **Side Panel** to the right of the view. Below is more information on these view options.

Under **Translation** in the **Nucleotide info**, there is an extra checkbox: **Relative to top sequence**. Checking this box will make the reading frames for the translation align with the top sequence so that you can compare the effect of nucleotide differences on the protein level.

The options in the **Alignment info** relate to each column in the alignment:

- **Consensus.** Shows a consensus sequence at the bottom of the alignment. The consensus sequence is based on every single position in the alignment and reflects an artificial sequence which resembles the sequence information of the alignment, but only as one single sequence. If all sequences of the alignment is 100% identical the consensus sequence will be identical to all sequences found in the alignment. If the sequences of the alignment differ the consensus sequence will reflect the most common sequences in the alignment. Parameters for adjusting the consensus sequences are described below.
  - **Limit.** This option determines how conserved the sequences must be in order to agree on a consensus. Here you can also choose **IUPAC** which will display the ambiguity code when there are differences between the sequences. E.g. an alignment with **A** and a **G** at the same position will display an **R** in the consensus line if the **IUPAC** option is selected. (The IUPAC codes can be found in section H and G.)
  - **No gaps.** Checking this option will not show gaps in the consensus.
  - **Ambiguous symbol.** Select how ambiguities should be displayed in the consensus line (as **N**, **?**, *****, . or **-**). This option has now effect if **IUPAC** is selected in the **Limit** list above.

  The **Consensus Sequence** can be opened in a new view, simply by right-clicking the **Consensus Sequence** and click **Open Consensus in New View**.
- **Conservation.** Displays the level of conservation at each position in the alignment. The conservation shows the conservation of all sequence positions. The height of the bar, or the gradient of the color reflect how conserved that particular position is in the alignment. If one position is 100% conserved the bar will be shown in full height, and it is colored in the color specified at the right side of the gradient slider.

  - **Foreground color.** Colors the letters using a gradient, where the right side color is used for highly conserved positions and the left side color is used for positions that are less conserved.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Graph.** Displays the conservation level as a graph at the bottom of the alignment. The bar (default view) show the conservation of all sequence positions. The height of the graph reflects how conserved that particular position is in the alignment. If one position is 100% conserved the graph will be shown in full height. Learn how to export the data behind the graph in section 7.4.
    * **Height.** Specifies the height of the graph.
    * **Type.** The type of the graph.
      · **Line plot.** Displays the graph as a line plot.
      · **Bar plot.** Displays the graph as a bar plot.
      · **Colors.** Displays the graph as a color bar using a gradient like the foreground and background colors.
    * **Color box.** Specifies the color of the graph for line and bar plots, and specifies a gradient for colors.

- **Gap fraction.** Which fraction of the sequences in the alignment that have gaps. The gap fraction is only relevant if there are gaps in the alignment.
  - **Foreground color.** Colors the letter using a gradient, where the left side color is used if there are relatively few gaps, and the right side color is used if there are relatively many gaps.
  - **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
  - **Graph.** Displays the gap fraction as a graph at the bottom of the alignment (Learn how to export the data behind the graph in section 7.4).
    * **Height.** Specifies the height of the graph.
    * **Type.** The type of the graph.
      · **Line plot.** Displays the graph as a line plot.
      · **Bar plot.** Displays the graph as a line plot.
      · **Colors.** Displays the graph as a color bar using a gradient like the foreground and background colors.
    * **Color box.** Specifies the color of the graph for line and bar plots, and specifies a gradient for colors.

- **Color different residues.** Indicates differences in aligned residues.
  - **Foreground color.** Colors the letter.
  - **Background color.** Sets a background color of the residues.

- **Sequence logo.** A sequence logo displays the frequencies of residues at each position in an alignment. This is presented as the relative heights of letters, along with the degree of sequence conservation as the total height of a stack of letters, measured in bits of information. The vertical scale is in bits, with a maximum of 2 bits for nucleotides and approximately 4.32 bits for amino acid residues. See section 18.2.1 for more details.

- **Foreground color.** Color the residues using a gradient according to the information content of the alignment column. Low values indicate columns with high variability whereas high values indicate columns with similar residues.
- **Background color.** Sets a background color of the residues using a gradient in the same way as described above.
- **Logo.** Displays sequence logo at the bottom of the alignment.
  - **Height.** Specifies the height of the sequence logo graph.
  - **Color.** The sequence logo can be displayed in black or Rasmol colors. For protein alignments, a polarity color scheme is also available, where hydrophobic residues are shown in black color, hydrophilic residues as green, acidic residues as red and basic residues as blue.

18.2.1 Bioinformatics explained: Sequence logo

In the search for homologous sequences, researchers are often interested in conserved sites/residues or positions in a sequence which tend to differ a lot. Most researches use alignments (see Bioinformatics explained*: multiple alignments*) for visualization of homology on a given set of either DNA or protein sequences. In proteins, active sites in a given protein family are often highly conserved. Thus, in an alignment these positions (which are not necessarily located in proximity) are fully or nearly fully conserved. On the other hand, antigen binding sites in the F_{ab} unit of immunoglobulins tend to differ quite a lot, whereas the rest of the protein remains relatively unchanged.

In DNA, promoter sites or other DNA binding sites are highly conserved (see figure 18.8). This is also the case for repressor sites as seen for the Cro repressor of bacteriophage λ.

When aligning such sequences, regardless of whether they are highly variable or highly conserved at specific sites, it is very difficult to generate a consensus sequence which covers the actual variability of a given position. In order to better understand the information content or significance of certain positions, a sequence logo can be used. The sequence logo displays the information content of all positions in an alignment as residues or nucleotides stacked on top of each other (see figure 18.8). The sequence logo provides a far more detailed view of the entire alignment than a simple consensus sequence. Sequence logos can aid to identify protein binding sites on DNA sequences and can also aid to identify conserved residues in aligned domains of protein sequences and a wide range of other applications.

Each position of the alignment and consequently the sequence logo shows the sequence information in a computed score based on Shannon entropy [Schneider and Stephens, 1990]. The height of the individual letters represent the sequence information content in that particular position of the alignment.

A sequence logo is a much better visualization tool than a simple consensus sequence. An example hereof is an alignment where in one position a particular residue is found in 70% of the sequences. If a consensus sequence is used, it typically only displays the single residue with 70% coverage. In figure 18.8 an un-gapped alignment of 11 *E. coli* start codons including flanking regions are shown. In this example, a consensus sequence would only display ATG as the start codon in position 1, but when looking at the sequence logo it is seen that a GTG is also allowed as a start codon.

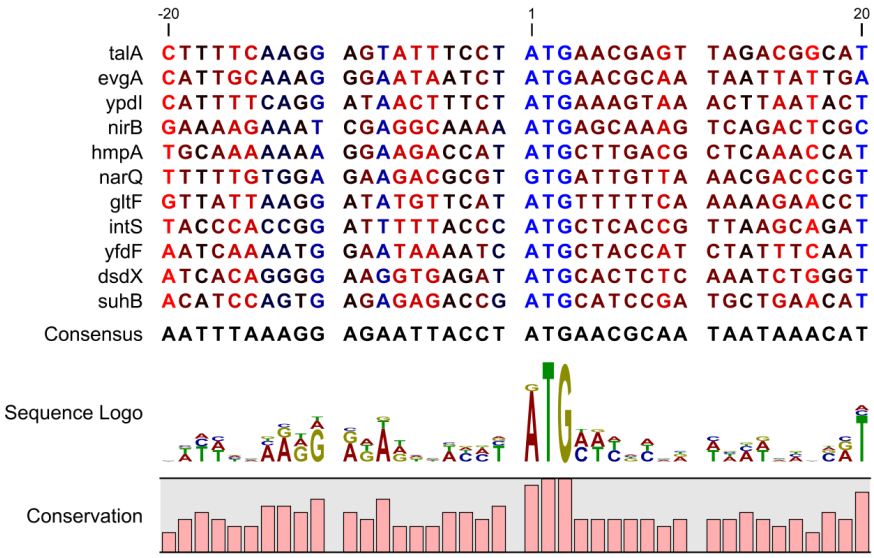


Figure 18.8: *Ungapped sequence alignment of eleven* E. coli *sequences defining a start codon. The start codons start at position 1. Below the alignment is shown the corresponding sequence logo. As seen, a GTG start codon and the usual ATG start codons are present in the alignment. This can also be visualized in the logo at position 1.*

Calculation of sequence logos

A comprehensive walk-through of the calculation of the information content in sequence logos is beyond the scope of this document but can be found in the original paper by [Schneider and Stephens, 1990]. Nevertheless, the conservation of every position is defined as R_{seq} which is the difference between the maximal entropy (S_{max}) and the observed entropy for the residue distribution (S_{obs}),

$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left(-\sum_{n=1}^{N} p_n \log_2 p_n \right)$$

p_n is the observed frequency of a amino acid residue or nucleotide of symbol n at a particular position and N is the number of distinct symbols for the sequence alphabet, either 20 for proteins or four for DNA/RNA. This means that the maximal sequence information content per position is $\log_2 4 = 2\ bits$ for DNA/RNA and $\log_2 20 \approx 4.32\ bits$ for proteins.

The original implementation by Schneider does not handle sequence gaps.

We have slightly modified the algorithm so an estimated logo is presented in areas with sequence gaps.

If amino acid residues or nucleotides of one sequence are found in an area containing gaps, we have chosen to show the particular residue as the fraction of the sequences. Example; if one position in the alignment contain 9 gaps and only one alanine (A) the A represented in the logo has a hight of 0.1.

Other useful resources
The website of Tom Schneider
http://www-lmmb.ncifcrf.gov/~toms/

WebLogo

http://weblogo.berkeley.edu/

[Crooks et al., 2004]

18.3 Edit alignments

18.3.1 Move residues and gaps

The placement of gaps in the alignment can be changed by modifying the parameters when creating the alignment (see section 18.1). However, gaps and residues can also be moved after the alignment is created:

select one or more gaps or residues in the alignment | drag the selection to move

This can be done both for single sequences, but also for multiple sequences by making a selection covering more than one sequence. When you have made the selection, the mouse pointer turns into a horizontal arrow indicating that the selection can be moved (see figure 18.9).

Note! Residues can only be moved when they are next to a gap.

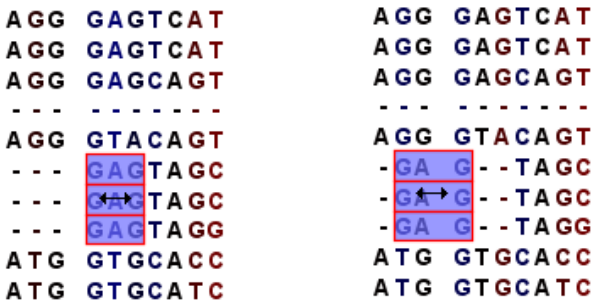


Figure 18.9: *Moving a part of an alignment. Notice the change of mouse pointer to a horizontal arrow.*

18.3.2 Insert gaps

The placement of gaps in the alignment can be changed by modifying the parameters when creating the alignment. However, gaps can also be added manually after the alignment is created.

To insert extra gaps:

select a part of the alignment | right-click the selection | Add gaps before/after

If you have made a selection covering e.g. five residues, a gap of five will be inserted. In this way you can easily control the number of gaps to insert. Gaps will be inserted in the sequences that you selected. If you make a selection in two sequences in an alignment, gaps will be inserted into these two sequences. This means that these two sequences will be displaced compared to the other sequences in the alignment.

18.3.3 Delete residues and gaps

Residues or gaps can be deleted for individual sequences or for the whole alignment. For individual sequences:

select the part of the sequence you want to delete | **right-click the selection** | **Edit Selection (** **)** | **Delete the text in the dialog** | **Replace**

The selection shown in the dialog will be replaced by the text you enter. If you delete the text, the selection will be replaced by an empty text, i.e. deleted.

To delete entire columns:

select the part of the alignment you want to delete | **right-click the selection** | **Delete columns**

The selection may cover one or more sequences, but the **Delete columns** function will always apply to the entire alignment.

18.3.4 Copy annotations to other sequences

Annotations on one sequence can be transferred to other sequences in the alignment:

right-click the annotation | **Copy Annotation to other Sequences**

This will display a dialog listing all the sequences in the alignment. Next to each sequence is a checkbox which is used for selecting which sequences, the annotation should be copied to. Click **Copy** to copy the annotation.

If you wish to copy all annotations on the sequence, click the **Copy All Annotations to other Sequences**.

18.3.5 Move sequences up and down

Sequences can be moved up and down in the alignment:

drag the name of the sequence up or down

When you move the mouse pointer over the label, the pointer will turn into a vertical arrow indicating that the sequence can be moved.

The sequences can also be sorted automatically to let you save time moving the sequences around. To sort the sequences alphabetically:

Right-click the name of a sequence | **Sort Sequences Alphabetically**

If you change the Sequence name (in the **Sequence Layout** view preferences), you will have to ask the program to sort the sequences again.

The sequences can also be sorted by similarity, grouping similar sequences together:

Right-click the name of a sequence | **Sort Sequences by Similarity**

18.3.6 Delete, rename and add sequences

Sequences can be removed from the alignment by right-clicking the label of a sequence:

right-click label | **Delete Sequence**

This can be undone by clicking **Undo** ( ) in the Toolbar.

If you wish to delete several sequences, you can check all the sequences, right-click and choose

Delete Marked Sequences. To show the checkboxes, you first have to click the **Show Selection Boxes** in the **Side Panel**.

A sequence can also be renamed:

right-click label | **Rename Sequence**

This will show a dialog, letting you rename the sequence. This will not affect the sequence that the alignment is based on.

Extra sequences can be added to the alignment by creating a new alignment where you select the current alignment and the extra sequences (see section 18.1).

The same procedure can be used for joining two alignments.

18.3.7 Realign selection

If you have created an alignment, it is possible to realign a part of it, leaving the rest of the alignment unchanged:

select a part of the alignment to realign | **right-click the selection** | **Realign selection**

This will open **Step 2** in the "Create alignment" dialog, allowing you to set the parameters for the realignment (see section 18.1).

It is possible for an alignment to become shorter or longer as a result of the realignment of a region. This is because gaps may have to be inserted in, or deleted from, the sequences not selected for realignment. This will only occur for entire columns of gaps in these sequences, ensuring that their relative alignment is unchanged.

Realigning a selection is a very powerful tool for editing alignments in several situations:

- **Removing changes.** If you change the alignment in a specific region by hand, you may end up being unhappy with the result. In this case you may of course undo your edits, but another option is to select the region and realign it.
- **Adjusting the number of gaps.** If you have a region in an alignment which has too many gaps in your opinion, you can select the region and realign it. By choosing a relatively high gap cost you will be able to reduce the number of gaps.
- **Combine with fixpoints.** If you have an alignment where two residues are not aligned, but you know that they should have been. You can now set an alignment fixpoint on each of the two residues, select the region and realign it using the fixpoints. Now, the two residues are aligned with each other and everything in the selected region around them is adjusted to accommodate this change.

18.4 Join alignments

CLC Protein Workbench can join several alignments into one. This feature can for example be used to construct "supergenes" for phylogenetic inference by joining alignments of several disjoint genes into one spliced alignment. Note, that when alignments are joined, all their annotations are carried over to the new spliced alignment.

Alignments can be joined by:

select alignments to join | **Toolbox in the Menu Bar** | **Alignments and Trees** ( ) | **Join Alignments** ( )

or **select alignments to join** | **right-click either selected alignment** | **Toolbox** | **Alignments and Trees** ( ) | **Join Alignments** ( )

This opens the dialog shown in figure 18.10.

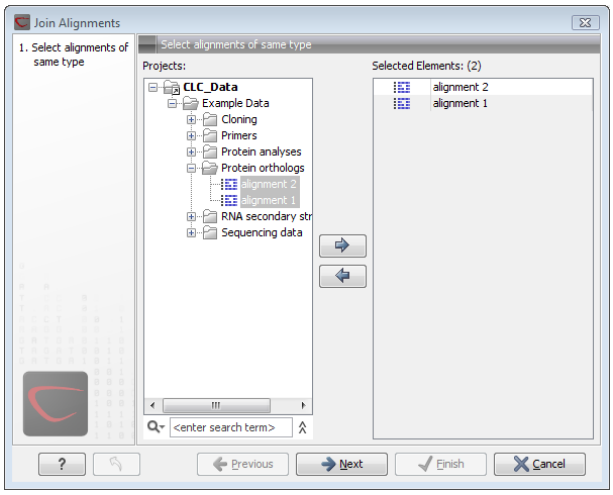


Figure 18.10: *Selecting two alignments to be joined.*

If you have selected some alignments before choosing the Toolbox action, they are now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove alignments from the selected elements. Click **Next** opens the dialog shown in figure 18.11.

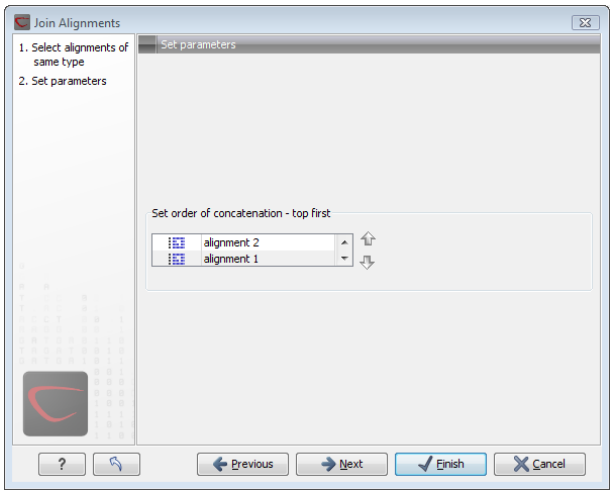


Figure 18.11: *Selecting order of concatenation.*

To adjust the order of concatenation, click the name of one of the alignments, and move it up or down using the arrow buttons.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**. The result is seen in figure 18.12.

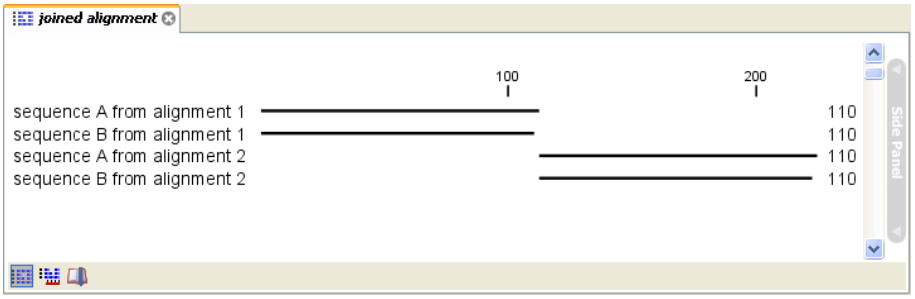


Figure 18.12: *The joining of the alignments result in one alignment containing rows of sequences corresponding to the number of uniquely named sequences in the joined alignments.*

18.4.1 How alignments are joined

Alignments are joined by considering the sequence names in the individual alignments. If two sequences from different alignments have identical names, they are considered to have the same origin and are thus joined. Consider the joining of alignments A and B. If a sequence named "in-A-and-B" is found in both A and B, the spliced alignment will contain a sequence named "in-A-and-B" which represents the characters from A and B joined in direct extension of each other. If a sequence with the name "in-A-not-B" is found in A but not in B, the spliced alignment will contain a sequence named "in-A-not-B". The first part of this sequence will contain the characters from A, but since no sequence information is available from B, a number of gap characters will be added to the end of the sequence corresponding to the number of residues in B. Note, that the function does not require that the individual alignments contain an equal number of sequences.

18.5 Pairwise comparison

For a given set of aligned sequences (see chapter 18) it is possible make a pairwise comparison in which each pair of sequences are compared to each other. This provides an overview of the diversity among the sequences in the alignment.

In *CLC Protein Workbench* this is done by creating a comparison table:

Toolbox in the Menu Bar | **Alignments and Trees** ( ) | **Pairwise Comparison** ( )

or **right-click alignment in Navigation Area** | **Toolbox** | **Alignments and Trees** ( ) | **Pairwise Comparison** ( )

This opens the dialog displayed in figure 18.13:

If an alignment was selected before choosing the Toolbox action, this alignment is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove elements from the **Navigation Area**. Click **Next** to adjust parameters.

18.5.1 Pairwise comparison on alignment selection

A pairwise comparison can also be performed for a selected part of an alignment:

right-click on an alignment selection | **Pairwise Comparison** ( )

This leads directly to the dialog described in the next section.

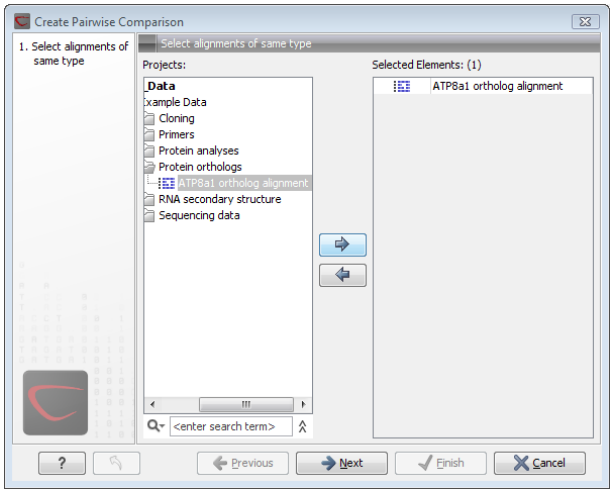


Figure 18.13: *Creating a pairwise comparison table.*

18.5.2 Pairwise comparison parameters

There are four kinds of comparison that can be made between the sequences in the alignment, as shown in figure 18.14.

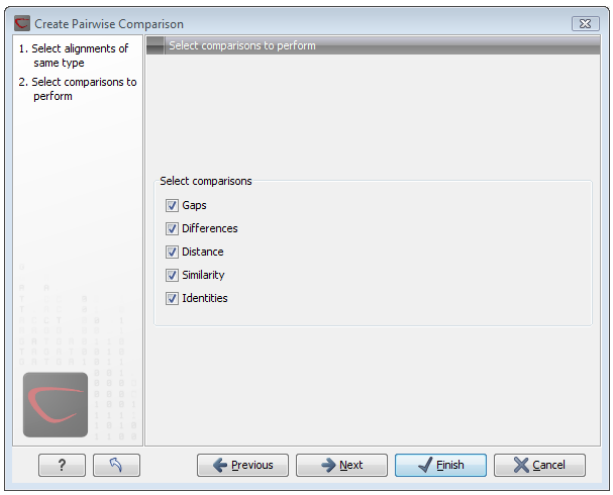


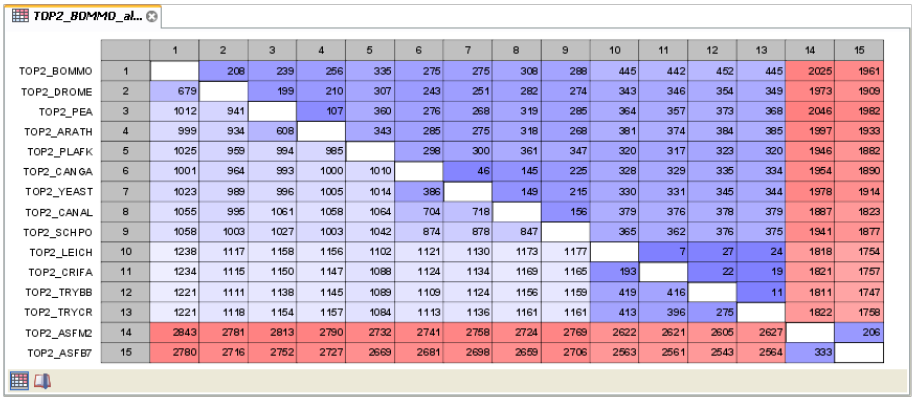
Figure 18.14: *Adjusting parameters for pairwise comparison.*

- **Gaps** Calculates the number of alignment positions where one sequence has a gap and the other does not.
- **Identities** Calculates the number of identical alignment positions to overlapping alignment positions between the two sequences.
- **Differences** Calculates the number of alignment positions where one sequence is different from the other. This includes gap differences as in the Gaps comparison.
- **Distance** Calculates the Jukes-Cantor distance between the two sequences. This number is given as the Jukes-Cantor correction of the proportion between identical and overlapping alignment positions between the two sequences.
- **Percent identity** Calculates the percentage of identical residues in alignment positions to overlapping alignment positions between the two sequences.

Click **Next** if you wish to adjust how to handle the results (see section 9.1). If not, click **Finish**.

18.5.3 The pairwise comparison table

The table shows the results of selected comparisons (see an example in figure 18.15). Since comparisons are often symmetric, the table can show the results of two comparisons at the same time, one in the upper-right and one in the lower-left triangle.



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
TOP2_BOMMO	1		208	239	256	335	275	275	308	288	445	442	452	445	2025	1961
TOP2_DROME	2	679		199	210	307	243	251	282	274	343	346	354	349	1973	1909
TOP2_PEA	3	1012	941		107	360	276	268	319	285	364	357	373	368	2046	1982
TOP2_ARATH	4	999	934	608		343	285	275	318	268	381	374	384	385	1997	1933
TOP2_PLAFK	5	1025	959	994	985		298	300	361	347	320	317	323	320	1946	1882
TOP2_CANGA	6	1001	964	993	1000	1010		46	145	225	328	329	335	334	1954	1890
TOP2_YEAST	7	1023	989	996	1005	1014	386		149	215	330	331	345	344	1978	1914
TOP2_CANAL	8	1055	995	1061	1058	1064	704	718		156	379	376	378	379	1887	1823
TOP2_SCHPO	9	1058	1003	1027	1003	1042	874	878	847		365	362	376	375	1941	1877
TOP2_LEICH	10	1238	1117	1158	1156	1102	1121	1130	1173	1177		7	27	24	1818	1754
TOP2_CRIFA	11	1234	1115	1150	1147	1088	1124	1134	1169	1165	193		22	19	1821	1757
TOP2_TRYBB	12	1221	1111	1138	1145	1089	1109	1124	1156	1159	419	416		11	1811	1747
TOP2_TRYCR	13	1221	1118	1154	1157	1084	1113	1136	1161	1161	413	396	275		1822	1758
TOP2_ASFM2	14	2843	2781	2813	2790	2732	2741	2758	2724	2769	2622	2621	2605	2627		206
TOP2_ASFB7	15	2780	2716	2752	2727	2669	2681	2698	2659	2706	2563	2561	2543	2564	333	

Figure 18.15: *A pairwise comparison table.*

The following settings are present in the side panel:

- **Contents**
  - **Upper comparison.** Selects the comparison to show in the upper triangle of the table
  - **Upper comparison gradient.** Selects the color gradient to use for the upper triangle.
  - **Lower comparison** Selects the comparison to show in the lower triangle. Choose the same comparison as in the upper triangle to show all the results of an asymmetric comparison.
  - **Lower comparison gradient.** Selects the color gradient to use for the lower triangle.
  - **Diagonal from upper.** Use this setting to show the diagonal results from the upper comparison.
  - **Diagonal from lower.** Use this setting to show the diagonal results from the lower comparison.
  - **No Diagonal.** Leaves the diagonal table entries blank.
- **Layout**
  - **Lock headers.** Locks the sequence labels and table headers when scrolling the table.
  - **Sequence label.** Changes the sequence labels.

- **Text format**
  - **Text size.** Changes the size of the table and the text within it.
  - **Font.** Changes the font in the table.
  - **Bold.** Toggles the use of boldface in the table.

18.6 Bioinformatics explained: Multiple alignments

Multiple alignments are at the core of bioinformatical analysis. Often the first step in a chain of bioinformatical analyses is to construct a multiple alignment of a number of homologs DNA or protein sequences. However, despite their frequent use, the development of multiple alignment algorithms remains one of the algorithmically most challenging areas in bioinformatical research.

Constructing a multiple alignment corresponds to developing a hypothesis of how a number of sequences have evolved through the processes of character substitution, insertion and deletion. The input to multiple alignment algorithms is a number of homologous sequences i.e. sequences that share a common ancestor and most often also share molecular function. The generated alignment is a table (see figure 18.16) where each row corresponds to an input sequence and each column corresponds to a position in the alignment. An individual column in this table represents residues that have all diverged from a common ancestral residue. Gaps in the table (commonly represented by a '-') represent positions where residues have been inserted or deleted and thus do not have ancestral counterparts in all sequences.

18.6.1 Use of multiple alignments

Once a multiple alignment is constructed it can form the basis for a number of analyses:

- The phylogenetic relationship of the sequences can be investigated by tree-building methods based on the alignment.
- Annotation of functional domains, which may only be known for a subset of the sequences, can be transferred to aligned positions in other un-annotated sequences.
- Conserved regions in the alignment can be found which are prime candidates for holding functionally important sites.
- Comparative bioinformatical analysis can be performed to identify functionally important regions.

18.6.2 Constructing multiple alignments

Whereas the optimal solution to the pairwise alignment problem can be found in reasonable time, the problem of constructing a multiple alignment is much harder.

The first major challenge in the multiple alignment procedure is how to rank different alignments i.e. which *scoring function* to use. Since the sequences have a shared history they are correlated through their *phylogeny* and the scoring function should ideally take this into account. Doing so is, however, not straightforward as it increases the number of model parameters considerably.

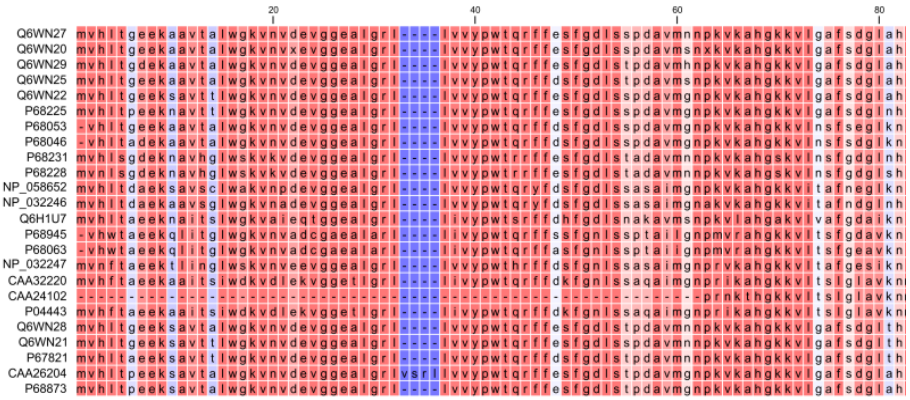


Figure 18.16: *The tabular format of a multiple alignment of 24 Hemoglobin protein sequences. Sequence names appear at the beginning of each row and the residue position is indicated by the numbers at the top of the alignment columns. The level of sequence conservation is shown on a color scale with blue residues being the least conserved and red residues being the most conserved.*

It is therefore commonplace to either ignore this complication and assume sequences to be unrelated, or to use heuristic corrections for shared ancestry.

The second challenge is to find the optimal alignment given a scoring function. For pairs of sequences this can be done by *dynamic programming* algorithms, but for more than three sequences this approach demands too much computer time and memory to be feasible.

A commonly used approach is therefore to do *progressive alignment* [Feng and Doolittle, 1987] where multiple alignments are built through the successive construction of pairwise alignments. These algorithms provide a good compromise between time spent and the quality of the resulting alignment

Presently, the most exciting development in multiple alignment methodology is the construction of *statistical alignment* algorithms [Hein, 2001], [Hein et al., 2000]. These algorithms employ a scoring function which incorporates the underlying phylogeny and use an explicit stochastic model of molecular evolution which makes it possible to compare different solutions in a statistically rigorous way. The optimization step, however, still relies on dynamic programming and practical use of these algorithms thus awaits further developments.

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

SOME RIGHTS RESERVED

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

Chapter 19

Phylogenetic trees

Contents

19.1 Inferring phylogenetic trees 301
19.1.1 Phylogenetic tree parameters 302
19.1.2 Tree View Preferences 304
19.2 Bioinformatics explained: phylogenetics 306
19.2.1 The phylogenetic tree 306
19.2.2 Modern usage of phylogenies 307
19.2.3 Reconstructing phylogenies from molecular data 307
19.2.4 Interpreting phylogenies 309

CLC Protein Workbench offers different ways of inferring phylogenetic trees. The first part of this chapter will briefly explain the different ways of inferring trees in *CLC Protein Workbench*. The second part, "Bioinformatics explained", will give a more general introduction to the concept of phylogeny and the associated bioinformatics methods.

19.1 Inferring phylogenetic trees

For a given set of aligned sequences (see chapter 18) it is possible to infer their evolutionary relationships. In *CLC Protein Workbench* this may be done either by using a distance based method (see "Bioinformatics explained" in section 19.2.) or by using the statistically founded maximum likelihood (ML) approach [Felsenstein, 1981]. Both approaches generate a phylogenetic tree. The tools are found in:

Toolbox | **Alignments and trees (** **)**

To generate a distance-based phylogenetic tree choose:

Create Tree (**)**

and to generate a maximum likelihood based phylogenetic tree choose:

Maximum Likelihood Phylogeny (**)**

In both cases the dialog displayed in figure 19.1 will be opened:

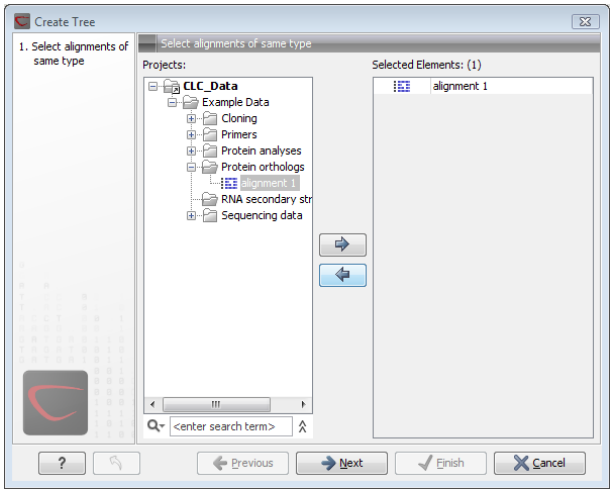


Figure 19.1: *Creating a Tree.*

If an alignment was selected before choosing the Toolbox action, this alignment is now listed in the **Selected Elements** window of the dialog. Use the arrows to add or remove elements from the **Navigation Area**. Click **Next** to adjust parameters.

19.1.1 Phylogenetic tree parameters

Distance-based methods

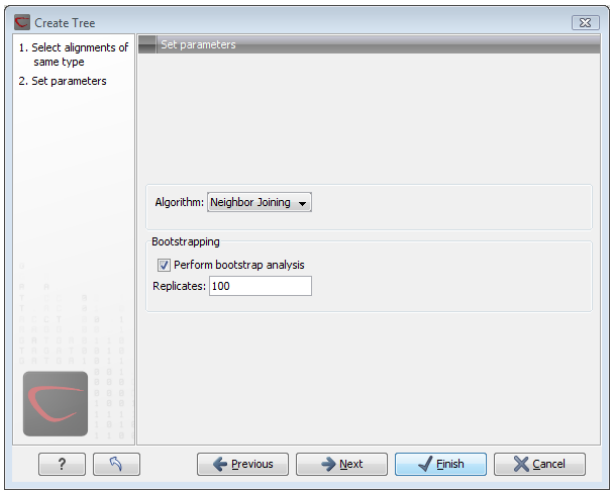


Figure 19.2: *Adjusting parameters for distance-based methods.*

Figure 19.2 shows the parameters that can be set for the distance-based methods:

- Algorithms
  - The **UPGMA** method assumes that evolution has occurred at a constant rate in the different lineages. This means that a root of the tree is also estimated.
  - The **neighbor joining** method builds a tree where the evolutionary rates are free to differ in different lineages. *CLC Protein Workbench* always draws trees with roots for practical reasons, but with the neighbor joining method, no particular biological hypothesis is postulated by the placement of the root. Figure 19.3 shows the difference between the two methods.

- To evaluate the reliability of the inferred trees, *CLC Protein Workbench* allows the option of doing a **bootstrap** analysis. A bootstrap value will be attached to each branch, and this value is a measure of the confidence in this branch. The number of replicates in the bootstrap analysis can be adjusted in the wizard. The default value is 100.

For a more detailed explanation, see "Bioinformatics explained" in section 19.2.

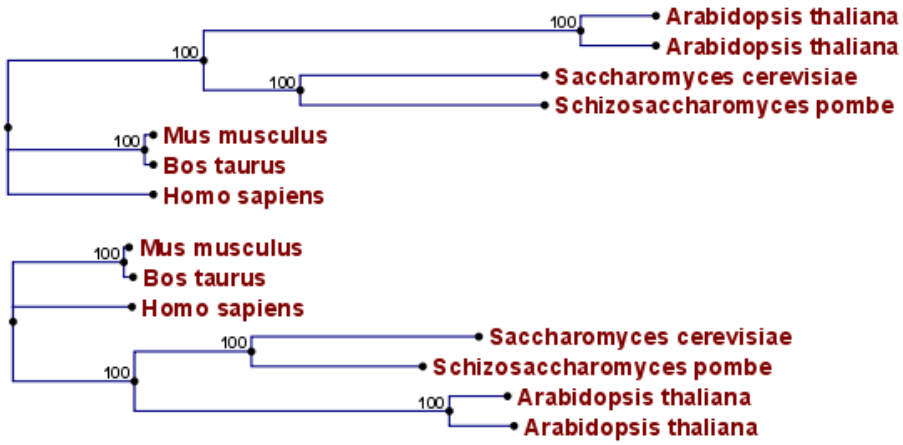


Figure 19.3: *Method choices for phylogenetic inference. The bottom shows a tree found by neighbor joining, while the top shows a tree found by UPGMA. The latter method assumes that the evolution occurs at a constant rate in different lineages.*

Maximum likelihood phylogeny

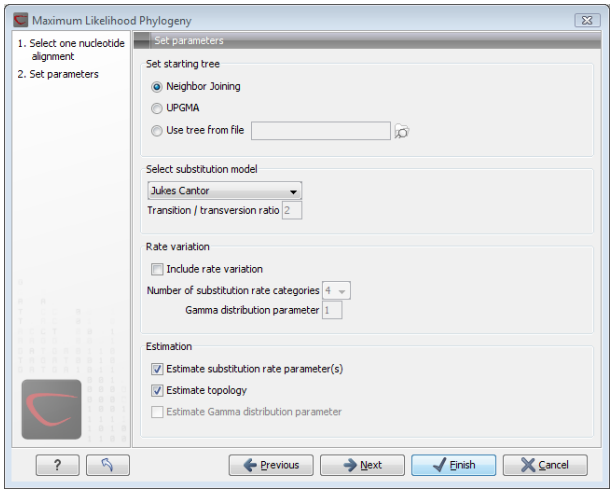


Figure 19.4: *Adjusting parameters for ML phylogeny*

Figure 19.4 shows the parameters that can be set for the ML phylogenetic tree reconstruction:

- **Starting tree**: the user is asked to specify a starting tree for the tree reconstruction. There are three possibilities
  - Neighbor joining
  - UPGMA

- Use tree from file.

- **Select substitution model**: *CLC Protein Workbench* allows maximum likelihood tree estimation to be performed under the assumption of one of four substitution models: the Jukes Cantor [Jukes and Cantor, 1969], the Kimura 80 [Kimura, 1980], the HKY [Hasegawa et al., 1985] and the GTR (also known as the REV model) [Yang, 1994a] models. All models are time-reversible. The JC and K80 models assume equal base frequencies and the HKY and GTR models allow the frequencies of the four bases to differ (they will be estimated by the observed frequencies of the bases in the alignment). In the JC model all substitutions are assumed to occur at equal rates, in the K80 and HKY models transition and transversion rates are allowed to differ. The GTR model is the general time reversible model and allows all substitutions to occur at different rates. In case of the K80 and HKY models the user may set a transtion/transversion ratio value which will be used as starting value or fixed, depending on the level of estimation chosen by the user (see below). For the substitution rate matrices describing the substitution models we use the parametrization of Yang [Yang, 1994a].

- **Rate variation**: in *CLC Protein Workbench* substitution rates may be allowed to differ among the individual nucleotide sites in the alignment by selecting the **include rate variation** box. When selected, the discrete gamma model of Yang [Yang, 1994b] is used to model rate variation among sites. The number of categories used in the dicretization of the gamma distribution as well as the gamma distribution parameter may be adjusted by the user (as the gamma distribution is restricted to have mean 1, there is only one parameter in the distribution)

- **Estimation** estimation is done according to the maximum likelihood principle, that is, a search is performed for the values of the free parameters in the model assumed that results in the highest likelihood of the observed alignment [Felsenstein, 1981]. By ticking the **estimate substitution rate parameters** box, maximum likelihood values of the free parameters in the rate matrix describing the assumed substitution model are found. If the **Estimate topology** box is selected, a search in the space of tree topologies for that which best explains the alignment is performed. If left un-ticked, the starting topology is kept fixed at that of the starting tree. The **Estimate Gamma distribution parameter** is active if rate variation has been included in the model and in this case allows estimation of the Gamma distribution parameter to be switched on or off. If the box is left un-ticked, the value is fixed at that given in the **Rate variation** part. In the absence of rate variation estimation of substitution parameters and branch lengths are carried out according to the expectation maximization algorithm [Dempster et al., 1977]. With rate variation the maximization algorithm is performed. The topology space is searched according to the PHYML method [Guindon and Gascuel, 2003], allowing efficient search and estimation of large phylogenies. Branch lengths are given in terms of expected numbers of substitutions per nucleotide site.

19.1.2 Tree View Preferences

The **Tree View** preferences are these:

- **Text format.** Changes the text format for all of the nodes the tree contains.

  - **Text size.** The size of the text representing the nodes can be modified in tiny, small, medium, large or huge.
  - **Font.** Sets the font of the text of all nodes
  - **Bold.** Sets the text bold if enabled.

- **Tree Layout.** Different layouts for the tree.
  - **Node symbol.** Changes the symbol of nodes into box, dot, circle or none if you don‘t want a node symbol.
  - **Layout.** Displays the tree layout as standard or topology.
  - **Show internal node labels**. This allows you to see labels for the internal nodes. Initially, there are no labels, but right-clicking a node allows you to type a label.
  - **Label color.** Changes the color of the labels on the tree nodes.
  - **Branch label color.** Modifies the color of the labels on the branches.
  - **Node color.** Sets the color of all nodes.
  - **Line color.** Alters the color of all lines in the tree.

- **Labels.** Specifies the text to be displayed in the tree.
  - **Nodes.** Sets the annotation of all nodes either to name or to species.
  - **Branches.** Changes the annotation of the branches to bootstrap, length or none if you don‘t want annotation on branches.

Note! Dragging in a tree will change it. You are therefore asked if you want to save this tree when the **Tree View** is closed.

You may select part of a **Tree** by clicking on the nodes that you want to select.

Right-click a selected node opens a menu with the following options:

- Set root above node (defines the root of the tree to be just above the selected node).
- Set root at this node (defines the root of the tree to be at the selected node).
- Toggle collapse (collapses or expands the branches below the node).
- Change label (allows you to label or to change the existing label of a node).
- Change branch label (allows you to change the existing label of a branch).

You can also relocate leaves and branches in a tree or change the length. It is possible to modify the text on the unit measurement at the bottom of the tree view by right-clicking the text. In this way you can specify a unit, e.g. "years".

Branch lengths are given in terms of expected numbers of substitutions per site.

Note! To drag branches of a tree, you must first click the node one time, and then click the node again, and this time hold the mouse button.

In order to change the representation:

- Rearrange leaves and branches by
  Select a leaf or branch | Move it up and down (Hint: The mouse turns into an arrow pointing up and down)
- Change the length of a branch by
  Select a leaf or branch | Press Ctrl | Move left and right (Hint: The mouse turns into an arrow pointing left and right)

Alter the preferences in the **Side Panel** for changing the presentation of the tree.

19.2 Bioinformatics explained: phylogenetics

Phylogenetics describes the taxonomical classification of organisms based on their evolutionary history i.e. their *phylogeny*. Phylogenetics is therefore an integral part of the science of *systematics* that aims to establish the phylogeny of organisms based on their characteristics. Furthermore, phylogenetics is central to evolutionary biology as a whole as it is the condensation of the overall paradigm of how life arose and developed on earth.

19.2.1 The phylogenetic tree

The evolutionary hypothesis of a phylogeny can be graphically represented by a phylogenetic tree.

Figure 19.5 shows a proposed phylogeny for the great apes, *Hominidae*, taken in part from Purvis [Purvis, 1995]. The tree consists of a number of nodes (also termed vertices) and branches (also termed edges). These nodes can represent either an individual, a species, or a higher grouping and are thus broadly termed taxonomical units. In this case, the terminal nodes (also called leaves or tips of the tree) represent extant species of *Hominidae* and are the *operational taxonomical units* (OTUs). The internal nodes, which here represent extinct common ancestors of the great apes, are termed *hypothetical taxonomical units* since they are not directly observable.

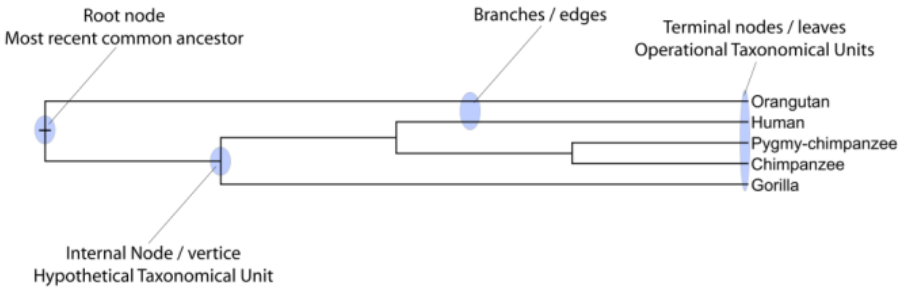


Figure 19.5: *A proposed phylogeny of the great apes (Hominidae). Different components of the tree are marked, see text for description.*

The ordering of the nodes determine the tree *topology* and describes how lineages have diverged over the course of evolution. The branches of the tree represent the amount of evolutionary divergence between two nodes in the tree and can be based on different measurements. A tree is completely specified by its topology and the set of all edge lengths.

The phylogenetic tree in figure 19.5 is rooted at the most recent common ancestor of all *Hominidae* species, and therefore represents a hypothesis of the direction of evolution e.g. that

the common ancestor of gorilla, chimpanzee and man existed before the common ancestor of chimpanzee and man. In contrast, an unrooted tree would represent relationships without assumptions about ancestry.

19.2.2 Modern usage of phylogenies

Besides evolutionary biology and systematics the inference of phylogenies is central to other areas of research.

As more and more genetic diversity is being revealed through the completion of multiple genomes, an active area of research within bioinformatics is the development of comparative machine learning algorithms that can simultaneously process data from multiple species [Siepel and Haussler, 2004]. Through the comparative approach, valuable evolutionary information can be obtained about which amino acid substitutions are functionally tolerant to the organism and which are not. This information can be used to identify substitutions that affect protein function and stability, and is of major importance to the study of proteins [Knudsen and Miyamoto, 2001]. Knowledge of the underlying phylogeny is, however, paramount to comparative methods of inference as the phylogeny describes the underlying correlation from shared history that exists between data from different species.

In molecular epidemiology of infectious diseases, phylogenetic inference is also an important tool. The very fast substitution rate of microorganisms, especially the RNA viruses, means that these show substantial genetic divergence over the time-scale of months and years. Therefore, the phylogenetic relationship between the pathogens from individuals in an epidemic can be resolved and contribute valuable epidemiological information about transmission chains and epidemiologically significant events [Leitner and Albert, 1999], [Forsberg et al., 2001].

19.2.3 Reconstructing phylogenies from molecular data

Traditionally, phylogenies have been constructed from morphological data, but following the growth of genetic information it has become common practice to construct phylogenies based on molecular data, known as *molecular phylogeny*. The data is most commonly represented in the form of DNA or protein sequences, but can also be in the form of e.g. restriction fragment length polymorphism (RFLP).

Methods for constructing molecular phylogenies can be distance based or character based.

Distance based methods
Two common algorithms, both based on pairwise distances, are the UPGMA and the Neighbor Joining algorithms. Thus, the first step in these analyses is to compute a matrix of pairwise distances between OTUs from their sequence differences. To correct for multiple substitutions it is common to use distances corrected by a model of molecular evolution such as the Jukes-Cantor model [Jukes and Cantor, 1969].

UPGMA. A simple but popular clustering algorithm for distance data is Unweighted Pair Group Method using Arithmetic averages (UPGMA) ([Michener and Sokal, 1957], [Sneath and Sokal, 1973]). This method works by initially having all sequences in separate clusters and continuously joining these. The tree is constructed by considering all initial clusters as leaf nodes in the tree, and each time two clusters are joined, a node is added to the tree as the parent of the two chosen nodes. The clusters to be joined are chosen as those with minimal pairwise distance. The branch lengths are set corresponding to the distance between clusters, which is calculated

as the average distance between pairs of sequences in each cluster.

The algorithm assumes that the distance data has the so-called *molecular clock* property i.e. the divergence of sequences occur at the same constant rate at all parts of the tree. This means that the leaves of UPGMA trees all line up at the extant sequences and that a root is estimated as part of the procedure.

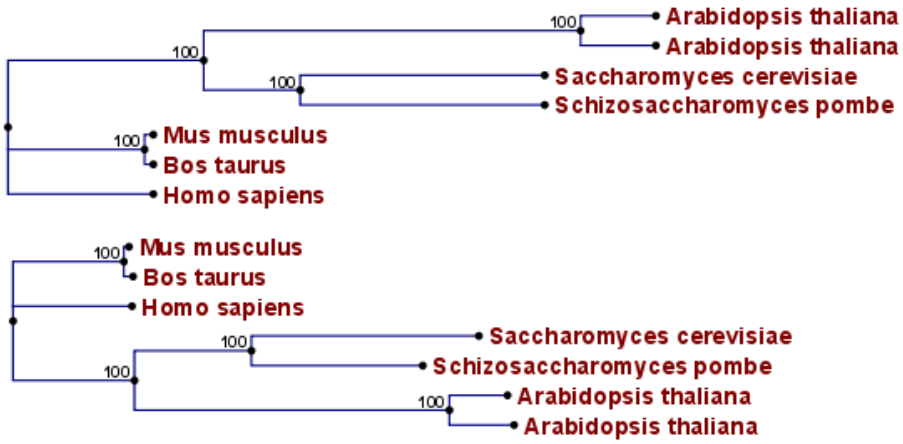


Figure 19.6: *Algorithm choices for phylogenetic inference. The bottom shows a tree found by the neighbor joining algorithm, while the top shows a tree found by the UPGMA algorithm. The latter algorithm assumes that the evolution occurs at a constant rate in different lineages.*

Neighbor Joining. The neighbor joining algorithm, [Saitou and Nei, 1987], on the other hand, builds a tree where the evolutionary rates are free to differ in different lineages, i.e., the tree does not have a particular root. Some programs always draw trees with roots for practical reasons, but for neighbor joining trees, no particular biological hypothesis is postulated by the placement of the root. The method works very much like UPGMA. The main difference is that instead of using pairwise distance, this method subtracts the distance to all other nodes from the pairwise distance. This is done to take care of situations where the two closest nodes are not neighbors in the "real" tree. The neighbor join algorithm is generally considered to be fairly good and is widely used. Algorithms that improves its cubic time performance exist. The improvement is only significant for quite large datasets.

Character based methods. Whereas the distance based methods compress all sequence information into a single number, the character based methods attempt to infer the phylogeny based on all the individual characters (nucleotides or amino acids).

Parsimony. In parsimony based methods a number of sites are defined which are informative about the topology of the tree. Based on these, the best topology is found by minimizing the number of substitutions needed to explain the informative sites. Parsimony methods are not based on explicit evolutionary models.

Maximum Likelihood. Maximum likelihood and Bayesian methods (see below) are probabilistic methods of inference. Both have the pleasing properties of using explicit models of molecular evolution and allowing for rigorous statistical inference. However, both approaches are very computer intensive.

A stochastic model of molecular evolution is used to assign a probability (likelihood) to each phylogeny, given the sequence data of the OTUs. Maximum likelihood inference [Felsenstein,

1981] then consists of finding the tree which assign the highest probability to the data.

Bayesian inference. The objective of Bayesian phylogenetic inference is not to infer a single "correct" phylogeny, but rather to obtain the full posterior probability distribution of all possible phylogenies. This is obtained by combining the likelihood and the prior probability distribution of evolutionary parameters. The vast number of possible trees means that bayesian phylogenetics must be performed by approximative Monte Carlo based methods. [Larget and Simon, 1999], [Yang and Rannala, 1997].

19.2.4 Interpreting phylogenies

Bootstrap values

A popular way of evaluating the reliability of an inferred phylogenetic tree is bootstrap analysis. The first step in a bootstrap analysis is to re-sample the alignment columns with replacement. I.e., in the re-sampled alignment, a given column in the original alignment may occur two or more times, while some columns may not be represented in the new alignment at all. The re-sampled alignment represents an estimate of how a different set of sequences from the same genes and the same species may have evolved on the same tree.

If a new tree reconstruction on the re-sampled alignment results in a tree similar to the original one, this increases the confidence in the original tree. If, on the other hand, the new tree looks very different, it means that the inferred tree is unreliable. By re-sampling a number of times it is possibly to put reliability weights on each internal branch of the inferred tree. If the data was bootstrapped a 100 times, a bootstrap score of 100 means that the corresponding branch occurs in all 100 trees made from re-sampled alignments. Thus, a high bootstrap score is a sign of greater reliability.

Other useful resources

The Tree of Life web-project
http://tolweb.org

Joseph Felsensteins list of phylogeny software
http://evolution.genetics.washington.edu/phylip/software.html

Creative Commons License

All CLC bio's scientific articles are licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 2.5 License. You are free to copy, distribute, display, and use the work for educational purposes, under the following conditions: You must attribute the work in its original form and "CLC bio" has to be clearly labeled as author and provider of the work. You may not use this work for commercial purposes. You may not alter, transform, nor build upon this work.

See http://creativecommons.org/licenses/by-nc-nd/2.5/ for more information on how to use the contents.

Part IV

Appendix

Appendix A

Comparison of workbenches

Below we list a number of functionalities that differ between CLC Workbenches and the CLC Sequence Viewer:

- CLC Sequence Viewer (■)
- CLC Protein Workbench (■)
- CLC DNA Workbench (■)
- CLC RNA Workbench (■)
- CLC Main Workbench (■)
- CLC Genomics Workbench (■)

Data handling	Viewer	Protein	DNA	RNA	Main	Genomics
Add multiple locations to Navigation Area		■	■	■	■	■
Share data on network drive		■	■	■	■	■
Search all your data		■	■	■	■	■

Assembly of sequencing data	Viewer	Protein	DNA	RNA	Main	Genomics
Advanced contig assembly			■		■	■
Importing and viewing trace data			■		■	■
Trim sequences			■		■	■
Assemble without use of reference sequence			■		■	■
Map to reference sequence			■		■	■
Assemble to existing contig			■		■	■
Viewing and edit contigs			■		■	■
Tabular view of an assembled contig (easy data overview)			■		■	■
Secondary peak calling			■		■	■
Multiplexing based on barcode or name			■		■	■

Next-generation Sequencing Data Analysis	Viewer	Protein	DNA	RNA	Main	Genomics
Import of 454, Illumina Genome Analyzer, SOLiD and Helicos data						■
Reference assembly of human-size genomes						■
De novo assembly						■
SNP/DIP detection						■
Graphical display of large contigs						■
Support for mixed-data assembly						■
Paired data support						■
RNA-Seq analysis						■
Expression profiling by tags						■
ChIP-Seq analysis						■

Expression Analysis	Viewer	Protein	DNA	RNA	Main	Genomics
Import of Illumina BeadChip, Affymetrix, GEO data					■	■
Import of Gene Ontology annotation files					■	■
Import of Custom expression data table and Custom annotation files					■	■
Multigroup comparisons					■	■
Advanced plots: scatter plot, volcano plot, box plot and MA plot					■	■
Hierarchical clustering					■	■
Statistical analysis on count-based and gaussian data					■	■
Annotation tests					■	■
Principal component analysis (PCA)					■	■
Hierarchical clustering and heat maps					■	■
Analysis of RNA-Seq/Tag profiling samples					■	■

Molecular cloning	Viewer	Protein	DNA	RNA	Main	Genomics
Advanced molecular cloning			■		■	■
Graphical display of in silico cloning			■		■	■
Advanced sequence manipulation			■		■	■

Database searches	Viewer	Protein	DNA	RNA	Main	Genomics
GenBank Entrez searches	■	■	■	■	■	■
UniProt searches (Swiss-Prot/TrEMBL)		■			■	■
Web-based sequence search using BLAST		■	■	■	■	■
BLAST on local database		■	■	■	■	■
Creation of local BLAST database		■	■	■	■	■
PubMed lookup		■	■	■	■	■
Web-based lookup of sequence data		■	■	■	■	■
Search for structures (at NCBI)		■			■	■

General sequence analyses	Viewer	Protein	DNA	RNA	Main	Genomics
Linear sequence view	■	■	■	■	■	■
Circular sequence view	■	■	■	■	■	■
Text based sequence view		■	■	■	■	■
Editing sequences	■	■	■	■	■	■
Adding and editing sequence annotations		■	■	■	■	■
Advanced annotation table		■	■	■	■	■
Join multiple sequences into one	■	■	■	■	■	■
Sequence statistics	■	■	■	■	■	■
Shuffle sequence	■	■	■	■	■	■
Local complexity region analyses		■	■	■	■	■
Advanced protein statistics		■			■	■
Comprehensive protein characteristics repor		■			■	■

Nucleotide analyses	Viewer	Protein	DNA	RNA	Main	Genomics
Basic gene finding	■	■	■	■	■	■
Reverse complement without loss of annota tion	■	■	■	■	■	■
Restriction site analysis	■	■	■	■	■	■
Advanced interactive restriction site analysis			■	■	■	■
Translation of sequences from DNA to pro- teins	■	■	■	■	■	■
Interactive translations of sequences and alignments		■	■	■	■	■
G/C content analyses and graphs		■	■	■	■	■

Protein analyses	Viewer	Protein	DNA	RNA	Main	Genomics
3D molecule view		■			■	■
Hydrophobicity analyses		■	■	■	■	■
Antigenicity analysis		■			■	■
Protein charge analysis		■			■	■
Reverse translation from protein to DNA		■	■	■	■	■
Proteolytic cleavage detection		■			■	■
Prediction of signal peptides (SignalP)		■			■	■
Transmembrane helix prediction (TMHMM)		■			■	■
Secondary protein structure prediction		■			■	■
PFAM domain search		■			■	■

Sequence alignment	Viewer	Protein	DNA	RNA	Main	Genomics
Multiple sequence alignments (Two algorithms)	■	■	■	■	■	■
Advanced re-alignment and fix-point alignment options		■	■	■	■	■
Advanced alignment editing options	■	■	■	■	■	■
Join multiple alignments into one		■	■	■	■	■
Consensus sequence determination and management	■	■	■	■	■	■
Conservation score along sequences	■	■	■	■	■	■
Sequence logo graphs along alignments		■	■	■	■	■
Gap fraction graphs		■	■	■	■	■
Copy annotations between sequences in alignments		■	■	■	■	■
Pairwise comparison	■	■	■	■	■	■

RNA secondary structure	Viewer	Protein	DNA	RNA	Main	Genomics
Advanced prediction of RNA secondary structure				■	■	■
Integrated use of base pairing constraints				■	■	■
Graphical view and editing of secondary structure				■	■	■
Info about energy contributions of structure elements				■	■	■
Prediction of multiple sub-optimal structures				■	■	■
Evaluate structure hypothesis				■	■	■
Structure scanning				■	■	■
Partition function				■	■	■

Dot plots	Viewer	Protein	DNA	RNA	Main	Genomics
Dot plot based analyses		■	■	■	■	■

Phylogenetic trees	Viewer	Protein	DNA	RNA	Main	Genomics
Neighbor-joining and UPGMA phylogenies	■	■	■	■	■	■
Maximum likelihood phylogeny of nucleotides		■	■	■	■	■

Pattern discovery	Viewer	Protein	DNA	RNA	Main	Genomics
Search for sequence match	■	■	■	■	■	■
Motif search for basic patterns		■	■	■	■	■
Motif search with regular expressions		■	■	■	■	■
Motif search with ProSite patterns		■	■	■	■	■
Pattern discovery		■	■	■	■	■

Primer design	Viewer	Protein	DNA	RNA	Main	Genomics
Advanced primer design tools			■		■	■
Detailed primer and probe parameters			■		■	■
Graphical display of primers			■		■	■
Generation of primer design output			■		■	■
Support for Standard PCR			■		■	■
Support for Nested PCR			■		■	■
Support for TaqMan PCR			■		■	■
Support for Sequencing primers			■		■	■
Alignment based primer design			■		■	■
Alignment based TaqMan probedesign			■		■	■
Match primer with sequence			■		■	■
Ordering of primers			■		■	■
Advanced analysis of primer properties			■		■	■

Molecular cloning	Viewer	Protein	DNA	RNA	Main	Genomics
Advanced molecular cloning			■		■	■
Graphical display of in silico cloning			■		■	■
Advanced sequence manipulation			■		■	■

Virtual gel view	Viewer	Protein	DNA	RNA	Main	Genomics
Fully integrated virtual 1D DNA gel simulator			■		■	■

For a more detailed comparison, we refer to http://www.clcbio.com/compare.

Appendix B

Graph preferences

This section explains the view settings of graphs. The **Graph preferences** at the top of the **Side Panel** includes the following settings:

- **Lock axes.** This will always show the axes even though the plot is zoomed to a detailed level.
- **Frame.** Shows a frame around the graph.
- **Show legends.** Shows the data legends.
- **Tick type**. Determine whether tick lines should be shown outside or inside the frame.
  - Outside
  - Inside
- **Tick lines at.** Choosing Major ticks will show a grid behind the graph.
  - None
  - Major ticks
- **Horizontal axis range**. Sets the range of the horizontal axis (x axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **Vertical axis range**. Sets the range of the vertical axis (y axis). Enter a value in **Min** and **Max**, and press Enter. This will update the view. If you wait a few seconds without pressing Enter, the view will also be updated.
- **X-axis at zero**. This will draw the x axis at y = 0. Note that the axis range will not be changed.
- **Y-axis at zero**. This will draw the y axis at x = 0. Note that the axis range will not be changed.
- **Show as histogram**. For some data-series it is possible to see the graph as a histogram rather than a line plot.

The **Lines and plots** below contains the following settings:

- **Dot type**
  - None
  - Cross
  - Plus
  - Square
  - Diamond
  - Circle
  - Triangle
  - Reverse triangle
  - Dot
- **Dot color.** Allows you to choose between many different colors. Click the color box to select a color.
- **Line width**
  - Thin
  - Medium
  - Wide
- **Line type**
  - None
  - Line
  - Long dash
  - Short dash
- **Line color.** Allows you to choose between many different colors. Click the color box to select a color.

For graphs with multiple data series, you can select which curve the dot and line preferences should apply to. This setting is at the top of the **Side Panel** group.

Note that the graph title and the axes titles can be edited simply by clicking with the mouse. These changes will be saved when you **Save** ( ) the graph - whereas the changes in the **Side Panel** need to be saved explicitly (see section 5.5).

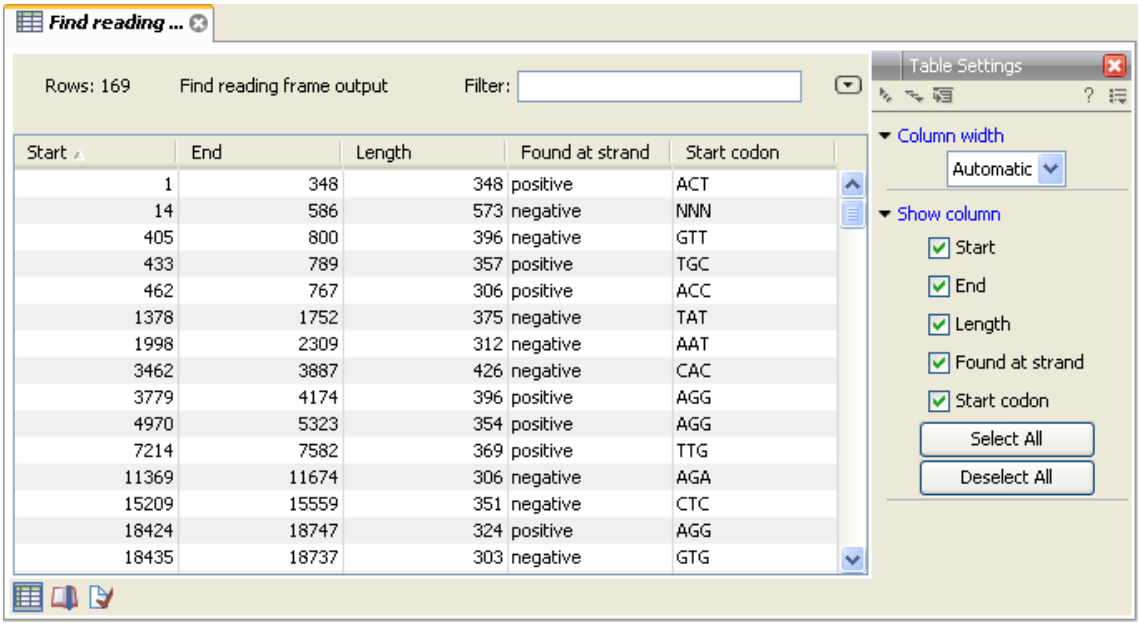
For more information about the graph view, please see section B.

Appendix C

Working with tables

Tables are used in a lot of places in the *CLC Protein Workbench*. The contents of the tables are of course different depending on the context, but there are some general features for all tables that will be explained in the following.

Figure C.1 shows an example of a typical table. This is the table result of **Find Open Reading Frames** (). We will use this table as an example in the following to illustrate the concepts that are relevant for all kinds of tables.



Start	End	Length	Found at strand	Start codon
1	348	348	positive	ACT
14	586	573	negative	NNN
405	800	396	negative	GTT
433	789	357	positive	TGC
462	767	306	positive	ACC
1378	1752	375	negative	TAT
1998	2309	312	negative	AAT
3462	3887	426	negative	CAC
3779	4174	396	positive	AGG
4970	5323	354	positive	AGG
7214	7582	369	positive	TTG
11369	11674	306	negative	AGA
15209	15559	351	negative	CTC
18424	18747	324	positive	AGG
18435	18737	303	negative	GTG

Figure C.1: *A table showing open reading frames.*

First of all, the columns of the table are listed in the **Side Panel** to the right of the table. By clicking the checkboxes you can hide/show the columns in the table.

Furthermore, you can **sort** the table by clicking on the column headers. (Pressing Ctrl - ⌘ on Mac - while you click will refine the existing sorting).

C.1 Filtering tables

The final concept to introduce is **Filtering**. The table filter as an advanced and a simple mode. The simple mode is the default and is applied simply by typing text or numbers (see an example in figure C.2).

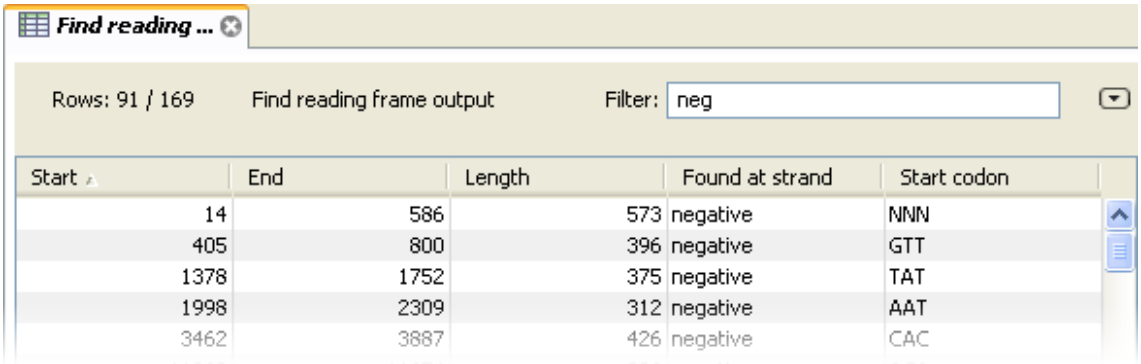


Figure C.2: *Typing "neg" in the filter in simple mode.*

Typing "neg" in the filter will only show the rows where "neg" is part of the text in any of the columns (also the ones that are not shown). The text does not have to be in the beginning, thus "ega" would give the same result. This simple filter works fine for fast, textual and non-complicated filtering and searching.

However, if you wish to make use of numerical information or make more complex filters, you can switch to the advanced mode by clicking the **Advanced filter** ( ) button. The advanced filter is structure in a different way: First of all, you can have more than one criterion in the filter. Criteria can be added or removed by clicking the **Add** ( ) or **Remove** ( ) buttons. At the top, you can choose whether all the criteria should be fulfilled (**Match all**), or if just one of the needs to be fulfilled (**Match any**).

For each filter criterion, you first have to select which column it should apply to. Next, you choose an operator. For numbers, you can choose between:

- **=** (equal to)
- **<** (smaller than)
- **>** (greater than)
- **<>** (not equal to)
- **abs. value <** (absolute value smaller than. This is useful if it doesn't matter whether the number is negative or positive)
- **abs. value >** (absolute value greater than. This is useful if it doesn't matter whether the number is negative or positive)

For text-based columns, you can choose between:

- **contains** (the text does not have to be in the beginning)
- **doesn't contain**

- = (the whole text in the table cell has to match, also lower/upper case)

Once you have chosen an operator, you can enter the text or numerical value to use.

If you wish to reset the filter, simply remove ( ) all the search criteria. Note that the last one will not disappear - it will be reset and allow you to start over.

Figure C.3 shows an example of an advanced filter which displays the open reading frames larger than 400 that are placed on the negative strand.

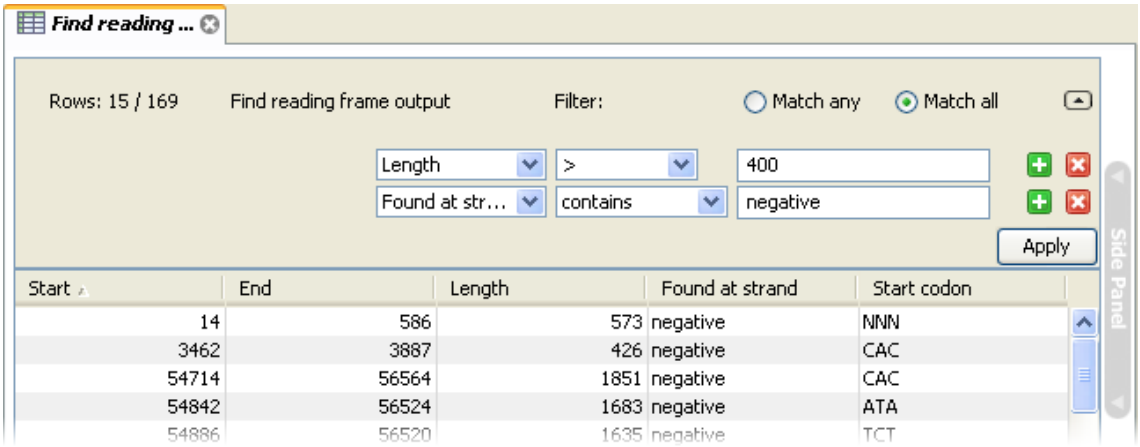


Figure C.3: *The advanced filter showing open reading frames larger than 400 that are placed on the negative strand.*

Both for the simple and the advanced filter, there is a counter at the upper left corner which tells you the number of rows that pass the filter (91 in figure C.2 and 15 in figure C.3).

Appendix D

BLAST databases

Several databases are available at NCBI, which can be selected to narrow down the possible BLAST hits.

D.1 Peptide sequence databases

- **nr.** Non-redundant GenBank CDS translations + PDB + SwissProt + PIR + PRF, excluding those in env_nr.
- **refseq.** Protein sequences from NCBI Reference Sequence project http://www.ncbi.nlm.nih.gov/RefSeq/.
- **swissprot.** Last major release of the SWISS-PROT protein sequence database (no incremental updates).
- **pat.** Proteins from the Patent division of GenBank.
- **pdb.** Sequences derived from the 3-dimensional structure records from the Protein Data Bank http://www.rcsb.org/pdb/.
- **env_nr.** Non-redundant CDS translations from env_nt entries.
- **month.** All new or revised GenBank CDS translations + PDB + SwissProt + PIR + PRF released in the last 30 days..

D.2 Nucleotide sequence databases

- **nr.** All GenBank + EMBL + DDBJ + PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant" due to computational cost.
- **refseq_rna.** mRNA sequences from NCBI Reference Sequence Project.
- **refseq_genomic.** Genomic sequences from NCBI Reference Sequence Project.
- **est.** Database of GenBank + EMBL + DDBJ sequences from EST division.
- **est_human.** Human subset of est.

- **est_mouse.** Mouse subset of est.
- **est_others.** Subset of est other than human or mouse.
- **gss.** Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
- **htgs.** Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2. Finished, phase 3 HTG sequences are in nr.
- **pat.** Nucleotides from the Patent division of GenBank.
- **pdb.** Sequences derived from the 3-dimensional structure records from Protein Data Bank. They are NOT the coding sequences for the corresponding proteins found in the same PDB record.
- **month.** All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
- **alu.** Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. See "Alu alert" by Claverie and Makalowski, Nature 371: 752 (1994).
- **dbsts.** Database of Sequence Tag Site entries from the STS division of GenBank + EMBL + DDBJ.
- **chromosome.** Complete genomes and complete chromosomes from the NCBI Reference Sequence project. It overlaps with refseq_genomic.
- **wgs.** Assemblies of Whole Genome Shotgun sequences.
- **env_nt.** Sequences from environmental samples, such as uncultured bacterial samples isolated from soil or marine samples. The largest single source is Sagarsso Sea project. This does overlap with nucleotide nr.

D.3 Adding more databases

Besides the databases that are part of the default configuration, you can add more databases located at NCBI by configuring files in the Workbench installation directory.

The list of databases that can be added is here: `http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/remote_blastdblist.html`.

In order to add a new database, find the `settings` folder in the Workbench installation directory (e.g. `C:\Program files\CLC Genomics Workbench 4`). Download unzip and place the following files in this directory to replace the built-in list of databases:

- Nucleotide databases: `http://www.clcbio.com/wbsettings/NCBI_BlastNucleotideDataba zip`
- Protein databases: `http://www.clcbio.com/wbsettings/NCBI_BlastProteinDatabases. zip`

Open the file you have downloaded into the`settings` folder, e.g. `NCBI_BlastProteinDatabases.proper` in a text editor and you will see the contents look like this:

```
nr[clcdefault] = Non-redundant protein sequences
refseq_protein = Reference proteins
swissprot = Swiss-Prot protein sequences
pat = Patented protein sequences
pdb = Protein Data Bank proteins
env_nr = Environmental samples
month = New or revised GenBank sequences
```

Simply add another database as a new line with the first item being the database name taken from http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/remote_blastdblist.html and the second part is the name to display in the Workbench. Restart the Workbench, and the new database will be visible in the BLAST dialog.

Appendix E

Proteolytic cleavage enzymes

Most proteolytic enzymes cleave at distinct patterns. Below is a compiled list of proteolytic enzymes used in *CLC Protein Workbench*.

Name	P4	P3	P2	P1	P1’	P2’
Cyanogen bromide (CNBr)	-	-	-	M	-	-
Asp-N endopeptidase	-	-	-	-	D	-
Arg-C	-	-	-	R	-	-
Lys-C	-	-	-	K	-	-
Trypsin	-	-	-	K, R	not P	-
Trypsin	-	-	W	K	P	-
Trypsin	-	-	M	R	P	-
Trypsin*	-	-	C, D	K	D	-
Trypsin*	-	-	C	K	H, Y	-
Trypsin*	-	-	C	R	K	-
Trypsin*	-	-	R	R	H,R	-
Chymotrypsin-high spec.	-	-	-	F, Y	not P	-
Chymotrypsin-high spec.	-	-	-	W	not M, P	-
Chymotrypsin-low spec.	-	-	-	F, L, Y	not P	-
Chymotrypsin-low spec.	-	-	-	W	not M, P	-
Chymotrypsin-low spec.	-	-	-	M	not P, Y	-
Chymotrypsin-low spec.	-	-	-	H	not D, M, P, W	-
o-Iodosobenzoate	-	-	-	W	-	-
Thermolysin	-	-	-	not D, E	A, F, I, L, M or V	-
Post-Pro	-	-	H, K, R	P	not P	-
Glu-C	-	-	-	E	-	-
Asp-N	-	-	-	-	D	-
Proteinase K	-	-	-	A, E, F, I, L, T, V, W, Y	-	-
Factor Xa	A, F, G, I, L, T, V, M	D,E	G	R	-	-
Granzyme B	I	E	P	D	-	-
Thrombin	-	-	G	R	G	-
Thrombin	A, F, G, I, L, T, V, M	A, F, G, I, L, T, V, W, A	P	R	not D, E	not D, E
TEV (Tobacco Etch Virus)	-	Y	-	Q	G, S	-

Appendix F

Formats for import and export

F.1 List of bioinformatic data formats

Below is a list of bioinformatic data formats, i.e. formats for importing and exporting sequences, alignments and trees.

F.1.1 Sequence data formats

File type	Suffix	Import	Export	Description
FASTA	.fsa/.fasta	X	X	Simple format, name & description
AB1	.ab1	X		Including chromatograms
ABI	.abi	X		Including chromatograms
CLC	.clc	X	X	Rich format including all information
Clone Manager	.cm5	X		
CSV export	.csv		X	Annotations in csv format
CSV import	.csv	X		One sequence per line: name; description(optional); sequence
DNAstrider	.str/.strider	X	X	
DS Gene	.bsml	X		
Embl	.embl	X	X	Only nucleotide sequence
GCG sequence	.gcg	X		Rich information incl. annotations
GenBank	.gbk/.gb/.gp	X	X	Rich information incl. annotations
Gene Construction Kit	.gck	X		
Lasergene	.pro/.seq	X		
Nexus	.nxs/.nexus	X	X	
Phred	.phd	X		Including chromatograms
PIR (NBRF)	.pir	X		Simple format, name & description
Raw sequence	any	X		Only sequence (no name)
SCF2	.scf	X		Including chromatograms
SCF3	.scf	X	X	Including chromatograms
Staden	.sdn	X		
Swiss-Prot	.swp	X	X	Rich information (only proteins)
Tab delimited text	.txt		X	Annotations in tab delimited text format
Vector NTI archives	.ma4/.pa4/.oa4	X		Archives in rich format
Vector NTI Database		X		Special import full database
Zip export	.zip		X	Selected files in CLC format
Zip import	.zip/.gzip./.tar	X		Contained files/folder structure

F.1.2 Contig formats

File type	Suffix	Import	Export	Description
ACE	.ace	X	X	No chromatogram or quality score
CLC	.clc	X	X	Rich format including all information
Zip export	.zip		X	Selected files in CLC format
Zip import	.zip/.gzip./.tar	X		Contained files/folder structure

F.1.3 Alignment formats

File type	Suffix	Import	Export	Description
CLC	.clc	X	X	Rich format including all information
Clustal Alignment	.aln	X	X	
GCG Alignment	.msf	X	X	
Nexus	.nxs/.nexus	X	X	
Phylip Alignment	.phy	X	X	
Zip export	.zip		X	Selected files in CLC format
Zip import	.zip/.gzip./.tar	X		Contained files/folder structure

F.1.4 Tree formats

File type	Suffix	Import	Export	Description
CLC	.clc	X	X	Rich format including all information
Newick	.nwk	X	X	
Nexus	.nxs/.nexus	X	X	
Zip export	.zip		X	Selected files in CLC format
Zip import	.zip/.gzip./.tar	X		Contained files/folder structure

F.1.5 Miscellaneous formats

File type	Suffix	Import	Export	Description
BLAST Database	.phr/.nhr	X		Link to database imported
CLC	.clc	X	X	Rich format including all information
CSV	.csv		X	All tables
Excel	.xls/.xlsx		X	All tables and reports
GFF	.gff	X	X	See http://www.clcbio.com/annotate-with-gff
mmCIF	.cif	X		3D structure
PDB	.pdb	X		3D structure
Tab delimited	.txt		X	All tables
Text	.txt	X	X	All data in a textual format
Zip export	.zip		X	Selected files in CLC format
Zip import	.zip/.gzip./.tar	X		Contained files/folder structure

Note! The Workbench can import 'external' files, too. This means that all kinds of files can be imported and displayed in the **Navigation Area**, but the above mentioned formats are the only ones whose *contents* can be shown in the Workbench.

F.2 List of graphics data formats

Below is a list of formats for exporting graphics. All data displayed in a graphical format can be exported using these formats. Data represented in lists and tables can only be exported in .pdf format (see section 7.3 for further details).

Format	Suffix	Type
Portable Network Graphics	.png	bitmap
JPEG	.jpg	bitmap
Tagged Image File	.tif	bitmap
PostScript	.ps	vector graphics
Encapsulated PostScript	.eps	vector graphics
Portable Document Format	.pdf	vector graphics
Scalable Vector Graphics	.svg	vector graphics

Appendix G

IUPAC codes for amino acids

(Single-letter codes based on International Union of Pure and Applied Chemistry)

The information is gathered from: http://www.ebi.ac.uk/2can/tutorials/aa.html

One-letter abbreviation	Three-letter abbreviation	Description
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
J	Xle	Leucine or Isoleucineucine
L	Leu	Leucine
I	ILe	Isoleucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
O	Pyl	Pyrrolysine
U	Sec	Selenocysteine
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Aspartic acid or Asparagine Asparagine
Z	Glx	Glutamic acid or Glutamine Glutamine
X	Xaa	Any amino acid

Appendix H

IUPAC codes for nucleotides

(Single-letter codes based on International Union of Pure and Applied Chemistry)

The information is gathered from: http://www.iupac.org and http://www.ebi.ac.uk/2can/tutorials/aa.html.

Code	Description
A	Adenine
C	Cytosine
G	Guanine
T	Thymine
U	Uracil
R	Purine (A or G)
Y	Pyrimidine (C, T, or U)
M	C or A
K	T, U, or G
W	T, U, or A
S	C or G
B	C, T, U, or G (not A)
D	A, T, U, or G (not C)
H	A, T, U, or C (not G)
V	A, C, or G (not T, not U)
N	Any base (A, C, G, T, or U)

Appendix I

Custom codon frequency tables

You can edit the list of codon frequency tables used by *CLC Protein Workbench*.

Note! Please be aware that this process needs to be handled carefully, otherwise you may have to re-install the Workbench to get it to work.

In the Workbench installation folder under `res`, there is a folder named `codonfreq`. This folder contains all the codon frequency tables organized into subfolders in a hierarchy. In order to change the tables, you simply add, delete or rename folders and the files in the folders. If you wish to add new tables, please use the existing ones as template.

Restart the Workbench to have the changes take effect.

Please note that when updating the Workbench to a new version, this information is not preserved. This means that you should keep this information in a separate place as back-up. (The ability to change the tables is mainly aimed at centrally deployed installations of the Workbench).

Bibliography

[Altschul and Gish, 1996] Altschul, S. F. and Gish, W. (1996). Local alignment statistics. *Methods Enzymol*, 266:460–480.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[Andrade et al., 1998] Andrade, M. A., O'Donoghue, S. I., and Rost, B. (1998). Adaptation of protein surfaces to subcellular location. *J Mol Biol*, 276(2):517–525.

[Bachmair et al., 1986] Bachmair, A., Finley, D., and Varshavsky, A. (1986). In vivo half-life of a protein is a function of its amino-terminal residue. *Science*, 234(4773):179–186.

[Bateman et al., 2004] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic Acids Res*, 32(Database issue):D138–D141.

[Bendtsen et al., 2004a] Bendtsen, J. D., Jensen, L. J., Blom, N., Heijne, G. V., and Brunak, S. (2004a). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel*, 17(4):349–356.

[Bendtsen et al., 2005] Bendtsen, J. D., Kiemer, L., Fausbøll, A., and Brunak, S. (2005). Non-classical protein secretion in bacteria. *BMC Microbiol*, 5:58.

[Bendtsen et al., 2004b] Bendtsen, J. D., Nielsen, H., von Heijne, G., and Brunak, S. (2004b). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, 340(4):783–795.

[Blobel, 2000] Blobel, G. (2000). Protein targeting (Nobel lecture). *Chembiochem.*, 1:86–102.

[Clote et al., 2005] Clote, P., Ferré, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591.

[Cornette et al., 1987] Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. (1987). Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol*, 195(3):659–685.

[Crooks et al., 2004] Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res*, 14(6):1188–1190.

[Dayhoff and Schwartz, 1978] Dayhoff, M. O. and Schwartz, R. M. (1978). *Atlas of Protein Sequence and Structure*, volume 3 of *5 suppl.*, pages 353–358. Nat. Biomed. Res. Found., Washington D.C.

[Dempster et al., 1977] Dempster, A., Laird, N., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38.

[Eddy, 2004] Eddy, S. R. (2004). Where did the BLOSUM62 alignment score matrix come from? *Nat Biotechnol*, 22(8):1035–1036.

[Eisenberg et al., 1984] Eisenberg, D., Schwarz, E., Komaromy, M., and Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J Mol Biol*, 179(1):125–142.

[Emini et al., 1985] Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985). Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *J Virol*, 55(3):836–839.

[Engelman et al., 1986] Engelman, D. M., Steitz, T. A., and Goldman, A. (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu Rev Biophys Biophys Chem*, 15:321–353.

[Felsenstein, 1981] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376.

[Feng and Doolittle, 1987] Feng, D. F. and Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J Mol Evol*, 25(4):351–360.

[Forsberg et al., 2001] Forsberg, R., Oleksiewicz, M. B., Petersen, A. M., Hein, J., Bøtner, A., and Storgaard, T. (2001). A molecular clock dates the common ancestor of European-type porcine reproductive and respiratory syndrome virus at more than 10 years before the emergence of disease. *Virology*, 289(2):174–179.

[Galperin and Koonin, 1998] Galperin, M. Y. and Koonin, E. V. (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol*, 1(1):55–67.

[Gill and von Hippel, 1989] Gill, S. C. and von Hippel, P. H. (1989). Calculation of protein extinction coefficients from amino acid sequence data. *Anal Biochem*, 182(2):319–326.

[Gonda et al., 1989] Gonda, D. K., Bachmair, A., Wünning, I., Tobias, J. W., Lane, W. S., and Varshavsky, A. (1989). Universality and structure of the N-end rule. *J Biol Chem*, 264(28):16700–16712.

[Guindon and Gascuel, 2003] Guindon, S. and Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, 52(5):696–704.

[Hasegawa et al., 1985] Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174.

[Hein, 2001] Hein, J. (2001). An algorithm for statistical alignment of sequences related by a binary tree. In *Pacific Symposium on Biocomputing*, page 179.

[Hein et al., 2000] Hein, J., Wiuf, C., Knudsen, B., Møller, M. B., and Wibling, G. (2000). Statistical alignment: computational properties, homology testing and goodness-of-fit. *J Mol Biol*, 302(1):265–279.

[Henikoff and Henikoff, 1992] Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):10915–10919.

[Hopp and Woods, 1983] Hopp, T. P. and Woods, K. R. (1983). A computer program for predicting protein antigenic determinants. *Mol Immunol*, 20(4):483–489.

[Ikai, 1980] Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *J Biochem (Tokyo)*, 88(6):1895–1898.

[Janin, 1979] Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature*, 277(5696):491–492.

[Jukes and Cantor, 1969] Jukes, T. and Cantor, C. (1969). *Mammalian Protein Metabolism*, chapter Evolution of protein molecules, pages 21–32. New York: Academic Press.

[Karplus and Schulz, 1985] Karplus, P. A. and Schulz, G. E. (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72:212–213.

[Kimura, 1980] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120.

[Klee and Ellis, 2005] Klee, E. W. and Ellis, L. B. M. (2005). Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, 6:256.

[Knudsen and Miyamoto, 2001] Knudsen, B. and Miyamoto, M. M. (2001). A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. *Proc Natl Acad Sci U S A*, 98(25):14512–14517.

[Kolaskar and Tongaonkar, 1990] Kolaskar, A. S. and Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett*, 276(1-2):172--174.

[Krogh et al., 2001] Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, 305(3):567–580.

[Kyte and Doolittle, 1982] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, 157(1):105–132.

[Larget and Simon, 1999] Larget, B. and Simon, D. (1999). Markov chain monte carlo algorithms for the bayesian analysis of phylogenetic trees. *Mol Biol Evol*, 16:750–759.

[Leitner and Albert, 1999] Leitner, T. and Albert, J. (1999). The molecular clock of HIV-1 unveiled through analysis of a known transmission history. *Proc Natl Acad Sci U S A*, 96(19):10752–10757.

[Maizel and Lenk, 1981] Maizel, J. V. and Lenk, R. P. (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc Natl Acad Sci U S A*, 78(12):7665–7669.

[McGinnis and Madden, 2004] McGinnis, S. and Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32(Web Server issue):W20–W25.

[Menne et al., 2000] Menne, K. M., Hermjakob, H., and Apweiler, R. (2000). A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, 16(8):741–742.

[Michener and Sokal, 1957] Michener, C. and Sokal, R. (1957). A quantitative approach to a problem in classification. *Evolution*, 11:130–162.

[Nielsen et al., 1997] Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*, 10(1):1–6.

[Purvis, 1995] Purvis, A. (1995). A composite estimate of primate phylogeny. *Philos Trans R Soc Lond B Biol Sci*, 348(1326):405–421.

[Reinhardt and Hubbard, 1998] Reinhardt, A. and Hubbard, T. (1998). Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*, 26(9):2230–2236.

[Rose et al., 1985] Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H. (1985). Hydrophobicity of amino acid residues in globular proteins. *Science*, 229(4716):834–838.

[Rost, 2001] Rost, B. (2001). Review: protein secondary structure prediction continues to rise. *J Struct Biol*, 134(2-3):204–218.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4):406–425.

[Schechter and Berger, 1967] Schechter, I. and Berger, A. (1967). On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun*, 27(2):157–162.

[Schechter and Berger, 1968] Schechter, I. and Berger, A. (1968). On the active site of proteases. 3. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochem Biophys Res Commun*, 32(5):898–902.

[Schneider and Stephens, 1990] Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100.

[Siepel and Haussler, 2004] Siepel, A. and Haussler, D. (2004). Combining phylogenetic and hidden Markov models in biosequence analysis. *J Comput Biol*, 11(2-3):413–428.

[Smith and Waterman, 1981] Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–197.

[Sneath and Sokal, 1973] Sneath, P. and Sokal, R. (1973). *Numerical Taxonomy*. Freeman, San Francisco.

[Tobias et al., 1991] Tobias, J. W., Shrader, T. E., Rocap, G., and Varshavsky, A. (1991). The N-end rule in bacteria. *Science*, 254(5036):1374–1377.

[von Heijne, 1986] von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucl. Acids Res.*, 14:4683–4690.

[Welling et al., 1985] Welling, G. W., Weijer, W. J., van der Zee, R., and Welling-Wester, S. (1985). Prediction of sequential antigenic regions in proteins. *FEBS Lett*, 188(2):215–218.

[Wootton and Federhen, 1993] Wootton, J. C. and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers in Chemistry*, 17:149–163.

[Yang, 1994a] Yang, Z. (1994a). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution*, 39(1):105–111.

[Yang, 1994b] Yang, Z. (1994b). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, 39(3):306–314.

[Yang and Rannala, 1997] Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. *Mol Biol Evol*, 14(7):717–724.

Part V

Index

Index

3D molecule view, 186
    export graphics, 193
    navigate, 187
    output, 193
    rotate, 187
    zoom, 187
454 sequencing data, 311

AB1, file format, 327
Abbreviations
    amino acids, 330
ABI, file format, 327
About CLC Workbenches, 27
Accession number, display, 67
.ace, file format, 329
ACE, file format, 328
Add
    annotations, 137, 312
    sequences to alignment, 294
Adjust selection, 129
Advanced preferences, 93
Advanced search, 86
Algorithm
    alignment, 282
    neighbor joining, 308
    UPGMA, 307
Align
    alignments, 285
    protein sequences, tutorial, 54
    sequences, 313
Alignment, *see* Alignments
Alignments, 282, 313
    add sequences to, 294
    compare, 296
    create, 283
    edit, 292
    fast algorithm, 284
    join, 294
    multiple, Bioinformatics explained, 299
    remove sequences from, 293
    view, 288
    view annotations on, 133
Aliphatic index, 211
.aln, file format, 329
Alphabetical sorting of folders, 65
Ambiguities, reverse translation, 258
Amino acid composition, 213
Amino acids
    abbreviations, 330
    UIPAC codes, 330
Annotation
    select, 129
Annotation Layout, in Side Panel, 133
Annotation types
    define your own, 137
Annotation Types, in Side Panel, 133
Annotations
    add, 137
    copy to other sequences, 293
    edit, 137, 139
    in alignments, 293
    introduction to, 133
    links, 159
    overview of, 136
    show/hide, 133
    table of, 136
    types of, 133
    view on sequence, 133
    viewing, 133
Annotations, add links to, 139
Antigenicity, 242, 313
Append wildcard, search, 149, 152, 155
Arrange
    layout of sequence, 38
    views in View Area, 73
Assembly, 311
Atomic composition, 213
Audit, 90

Backup, 108
Batch edit element properties, 69
Batch processing

log of, 120
Bibliography, 338
Bioinformatic data
export, 107
formats, 102, 326
bl2seq, *see* Local BLAST
BLAST, 312
against a local Database, 165
against NCBI, 162
create database from file system, 174
create database from Navigation Area, 174
create local database, 174
database file format, 329
database management, 175
graphics output, 169
list of databases, 321
parameters, 163
search, 161, 162
specify server URL, 94
table output, 170
tips for specialized searches, 47
tutorial, 44, 47
URL, 94
BLAST database index, 174
BLAST DNA sequence
BLASTn, 162
BLASTx, 162
tBLASTx, 162
BLAST Protein sequence
BLASTp, 163
tBLASTn, 163
BLAST result
search in, 172
BLAST search
Bioinformatics explained, 176
BLOSUM, scoring matrices, 204
Bootstrap values, 309
Borrow floating license, 25
Browser,import sequence from, 104
Bug reporting, 28

C/G content, 126
CDS, translate to protein, 129
Chain flexibility, 126
Cheap end gaps, 284
ChIP-Seq analysis, 311
.cif, file format, 186, 329
Circular view of sequence, 130, 312
.clc, file format, 108, 329
CLC Standard Settings, 96
CLC Workbenches, 27
CLC, file format, 327–329
associating with *CLC Protein Workbench*, 13
Cleavage, 259
the Peptidase Database, 263
Clone Manager, file format, 327
Cloning, 312, 315
Close view, 71
Clustal, file format, 328
Coding sequence, translate to protein, 129
Codon
frequency tables, reverse translation, 257
usage, 258
.col, file format, 329
Color residues, 289
Comments, 141
Common name
batch edit, 69
Compare workbenches, 311
Complexity plot, 207
Configure network, 33
Conflicting enzymes, 274
Consensus sequence, 288, 313
open, 288
Conservation, 288
graphs, 313
Contact information, 12
Contig, 311
Copy, 115
annotations in alignments, 293
elements in Navigation Area, 65
into sequence, 130
search results, GenBank, 151
search results, structure search, 157
search results, UniProt, 154
sequence, 142, 143
sequence selection, 227
text selection, 142
.cpf, file format, 94
.chp, file format, 329
Create
alignment, 283
dot plots, 197
enzyme list, 278
local BLAST database, 174
new folder, 65
workspace, 79

Create index file, BLAST database, 174
CSV
    export graph data points, 113
    formatting of decimal numbers, 107
.csv, file format, 329
CSV, file format, 327, 329
.ct, file format, 329
Custom annotation types, 137

Data
    storage location, 63
Data formats
    bioinformatic, 326
    graphics, 329
Data sharing, 63
Data structure, 63
Database
    GenBank, 148
    local, 63
    NCBI, 173
    nucleotide, 321
    peptide, 321
    shared BLAST database, 173
    structure, 154
    UniProt, 152
Db source, 141
db_xref references, 159
Delete
    element, 68
    residues and gaps in alignment, 292
    workspace, 80
Description, 141
    batch edit, 69
DGE, 312
Digital gene expression, 312
DIP detection, 311
Dipeptide distribution, 214
Discovery studio
    file format, 327
Distance, pairwise comparison of sequences in alignments, 298
DNA translation, 228
DNAstrider, file format, 327
Dot plots, 314
    Bioinformatics explained, 200
    create, 197
    print, 199
Double cutters, 267
Double stranded DNA, 123
Download and open
    search results, GenBank, 151, 157
    search results, UniProt, 154
Download and save
    search results, GenBank, 151, 157
    search results, UniProt, 154
Download of *CLC Protein Workbench*, 12
Drag and drop
    Navigation Area, 65
    search results, GenBank, 151, 157
    search results, UniProt, 154
DS Gene
    file format, 327

Edit
    alignments, 292, 313
    annotations, 137, 139, 312
    enzymes, 268
    sequence, 130
    sequences, 312
    single bases, 130
Element
    delete, 68
    rename, 68
.embl, file format, 329
Embl, file format, 327
Encapsulated PostScript, export, 111
End gap cost, 284
End gap costs
    cheap end caps, 284
    free end gaps, 284
Enzyme list, 278
    create, 278
    edit, 280
    view, 280
.eps-format, export, 111
Error reports, 28
Example data, import, 30
Excel, export file format, 329
Expand selection, 129
Expect, BLAST search, 170
Export
    bioinformatic data, 107
    dependent objects, 108
    folder, 107
    graph in csv format, 113
    graphics, 109
    history, 108
    list of formats, 326

multiple files, 107
preferences, 94
Side Panel Settings, 92
tables, 329
Export visible area, 110
Export whole view, 110
Expression analysis, 312
Extensions, 30
External files, import and export, 109
Extinction coefficient, 212
Extract sequences, 145

FASTA, file format, 327
Feature request, 28
Feature table, 214
Features, *see* Annotations
File system, local BLAST database, 174
Filtering restriction enzymes, 269, 271, 279
Find
in GenBank file, 142
in sequence, 127
results from a finished process, 78
Find open reading frames, 230
Fit to pages, print, 100
Fit Width, 77
Fixpoints, for alignments, 286
Floating license, 24
Floating license: use offline, 25
Floating Side Panel, 97
Folder, create new, tutorial, 38
Follow selection, 123
Footer, 101
Format, of the manual, 35
FormatDB, 174
Fragment table, 274
Fragment, select, 129
Fragments, separate on gel, 276
Free end gaps, 284
.fsa, file format, 329

G/C content, 126, 313
Gap
compare number of, 298
delete, 292
extension cost, 284
fraction, 289, 313
insert, 292
open cost, 284
Gb Division, 141
.gbk, file format, 329
GCG Alignment, file format, 328
GCG Sequence, file format, 327
.gck, file format, 329
GCK, Gene Construction Kit file format, 327
Gel
separate sequences without restriction enzyme digestion, 276
tabular view of fragments, 274
Gel electrophoresis, 275, 315
marker, 278
view, 276
view preferences, 276
when finding restriction sites, 273
GenBank
view sequence in, 142
file format, 327
search, 148, 312
search sequence in, 158
tutorial, 43
Gene Construction Kit, file format, 327
Gene expression analysis, 312
Gene finding, 230
General preferences, 89
General Sequence Analyses, 195
Genetic code, reverse translation, 257
Getting started tutorial, 37
.gff, file format, 329
Google sequence, 158
Graph
export data points in csv format, 113
Graph Side Panel, 316
Graphics
data formats, 329
export, 109
.gzip, file format, 329
Gzip, file format, 329

Half-life, 212
Handling of results, 118
Header, 101
Heat map, 312
Help, 29
Hide/show Toolbox, 79
High-throughput sequencing, 311
History, 116
export, 108
preserve when exporting, 117
source elements, 117

Homology, pairwise comparison of sequences in alignments, 298
Hydrophobicity, 244, 313
Bioinformatics explained, 247
Chain Flexibility, 248
Cornette, 126, 248
Eisenberg, 126, 247
Emini, 126
Engelman (GES), 126, 247
Hopp-Woods, 126, 248
Janin, 126, 248
Karplus and Schulz, 126
Kolaskar-Tongaonkar, 126, 248
Kyte-Doolittle, 126, 247
Rose, 248
Surface Probability, 248
Welling, 126, 248

ID, license, 19
Illumina Genome Analyzer, 311
Import
bioinformatic data, 103, 104
existing data, 38
FASTA-data, 38
from a web page, 104
list of formats, 326
preferences, 94
raw sequence, 104
Side Panel Settings, 92
using copy paste, 104
Index for searching, 88
Infer Phylogenetic Tree, 301
Insert
gaps, 292
Installation, 12
Invert sequence, 228
Isoelectric point, 211
IUPAC codes
nucleotides, 332

Join
alignments, 294
sequences, 214
.jpg-format, export, 111

Keywords, 141

Label
of sequence, 123
Landscape, Print orientation, 100
Lasergene sequence
file format, 327
Latin name
batch edit, 69
Length, 141
License, 15
ID, 19
starting without a license, 27
License server, 24
License server: access offline, 25
Limited mode, 27
Links, from annotations, 139
Linux
installation, 14
installation with RPM-package, 15
List of restriction enzymes, 278
List of sequences, 143
Load enzyme list, 268
Local BLAST, 165
Local BLAST Database, 174
Local BLAST database management, 175
Local BLAST Databases, 172
Local complexity plot, 207, 312
Local Database, BLAST, 165
Locale setting, 90
Location
search in, 86
of selection on sequence, 77
path to, 63
Side Panel, 91
Locations
multiple, 311
Log of batch processing, 120
Logo, sequence, 289, 313

.ma4, file format, 329
Mac OS X installation, 13
Manage BLAST databases, 175
Manipulate sequences, 312, 315
Manual editing, auditing, 90
Manual format, 34
Marker, in gel view, 278
Maximize size of view, 74
Maximum likelihood, 314
Menu Bar, illustration, 62
MFold, 314
mmCIF, file format, 329
Mode toolbar, 76
Modification date, 141

Modify enzyme list, 280
Modules, 30
Molecular weight, 211
Motif list, 223
Motif search, 217, 223, 314
Mouse modes, 76
Move
    content of a view, 77
    elements in Navigation Area, 65
    sequences in alignment, 293
.msf, file format, 329
Multiple alignments, 299, 313
Multiselecting, 65

Name, 141
Navigate, 3D structure, 187
Navigation Area, 62
    create local BLAST database, 174
    illustration, 62
NCBI, 148
    search for structures, 154
    search sequence in, 158
    search, tutorial, 43
NCBI BLAST
    add more databases, 322
Negatively charged residues, 213
Neighbor Joining algorithm, 308
Neighbor-joining, 314
Nested PCR primers, 314
Network configuration, 33
Network drive, shared BLAST database, 173
Never show this dialog again, 90
New
    feature request, 28
    folder, 65
    folder, tutorial, 38
    sequence, 142
New sequence
    create from a selection, 129
Newick, file format, 328
Next-Generation Sequencing, 311
.nexus, file format, 329
Nexus, file format, 327, 328
NGS, 311
.nhr, file format, 329
NHR, file format, 329
Non-standard residues, 125
Nucleotide
    info, 125
    sequence databases, 321
Nucleotides
    UIPAC codes, 332
Numbers on sequence, 123
.nwk, file format, 329
.nxs, file format, 329

.oa4, file format, 329
Open
    consensus sequence, 288
    from clipboard, 104
Open reading frame determination, 230
Open-ended sequence, 230
Order primers, 314
ORF, 230
Organism, 141
Origins from, 117
Overhang
    of fragments from restriction digest, 274
Overhang, find restriction enzymes based on, 269, 271, 279

.pa4, file format, 329
Page heading, 101
Page number, 101
Page setup, 100
Pairwise comparison, 296
PAM, scoring matrices, 204
Parameters
    search, 149, 152, 155
Partition function, 314
Paste
    text to create a new sequence, 104
Paste/copy, 115
Pattern Discovery, 215
Pattern discovery, 314
Pattern Search, 217
PCR primers, 314
.pdb, file format, 186, 329
.seq, file format, 329
PDB, file format, 329
.pdf-format, export, 111
Peptidase, 259
Peptide sequence databases, 321
Percent identity, pairwise comparison of sequences in alignments, 298
Personal information, 28
Pfam domain search, 249, 313
.phr, file format, 329

PHR, file format, 329
Phred, file format, 327
.phy, file format, 329
Phylip, file format, 328
Phylogenetic tree, 301, 314
    tutorial, 56
Phylogenetics, Bioinformatics explained, 306
.pir, file format, 329
PIR (NBRF), file format, 327
Plot
    dot plot, 197
    local complexity, 207
Plug-ins, 30
.png-format, export, 111
Polarity colors, 125
Portrait, Print orientation, 100
Positively charged residues, 213
PostScript, export, 111
Preference group, 95
Preferences, 89
    advanced, 93
    export, 94
    General, 89
    import, 94
    style sheet, 95
    toolbar, 91
    View, 90
    view, 75
Primer
    design, 314
    design from alignments, 314
Print, 98
    3D molecule view, 193
    dot plots, 199
    preview, 101
    visible area, 99
    whole view, 99
.pro, file format, 329
Problems when starting up, 29
Processes, 78
Properties, batch edit, 69
Protease, cleavage, 259
Protein
    charge, 240, 313
    cleavage, 259
    hydrophobicity, 247
    Isoelectric point, 211
    report, 253, 312
    report, output, 254
    signal peptide, 234
    statistics, 211
    structure prediction, 251
    translation, 255
Proteolytic cleavage, 259, 313
    Bioinformatics explained, 262
    tutorial, 52
Proteolytic enzymes cleavage patterns, 324
Proxy server, 33
.ps-format, export, 111
.psi, file format, 329
PubMed references, search, 159
PubMed references,search, 312

Quick start, 29

Rasmol colors, 125
Reading frame, 230
Realign alignment, 313
Rebase, restriction enzyme database, 278
Rebuild index, 88
Recycle Bin, 68
Redo alignment, 285
Redo/Undo, 72
Reference sequence, 311
References, 338
Region
    types, 130
Remove
    annotations, 140
    sequences from alignment, 293
    terminated processes, 78
Rename element, 68
Report program errors, 28
Report, protein, 312
Request new feature, 28
Residue coloring, 125
Restore
    deleted elements, 68
    size of view, 74
Restriction enzmyes
    filter, 269, 271, 279
    from certain suppliers, 269, 271, 279
Restriction enzyme list, 278
Restriction enzyme, star activity, 278
Restriction enzymes
    methylation, 269, 271, 279
    number of cut sites, 267

overhang, 269, 271, 279
separate on gel, 276
sorting, 267
Restriction sites, 313
enzyme database Rebase, 278
select fragment, 129
number of, 271
on sequence, 124, 266
parameters, 270
tutorial, 57
Results handling, 118
Reverse complement, 227, 313
Reverse sequence, 228
Reverse translation, 255, 313
Bioinformatics explained, 257
Right-click on Mac, 34
RNA secondary structure, 314
RNA translation, 228
RNA-Seq analysis, 311
.rnaml, file format, 329
Rotate, 3D structure, 187

Safe mode, 29
Save
changes in a view, 72
sequence, 44
style sheet, 95
view preferences, 95
workspace, 79
Save enzyme list, 268
SCF2, file format, 327
SCF3, file format, 327
Score, BLAST search, 170
Scoring matrices
Bioinformatics explained, 204
BLOSUM, 204
PAM, 204
Scroll wheel
to zoom in, 76
to zoom out, 76
Search, 86
in one location, 86
BLAST, 161, 162
for structures at NCBI, 154
GenBank, 148
GenBank file, 142
handle results from GenBank, 150
handle results from NCBI structure DB, 156
handle results from UniProt, 153
hits, number of, 90
in a sequence, 127
in annotations, 127
in Navigation Area, 84
Local BLAST, 165
local data, 311
options, GenBank, 149
options, GenBank structure search, 155
options, UniProt, 152
own motifs, 223
parameters, 149, 152, 155
patterns, 215, 217
Pfam domains, 249
PubMed references, 159
sequence in UniProt, 159
sequence on Google, 158
sequence on NCBI, 158
sequence on web, 158
TrEMBL, 152
troubleshooting, 88
UniProt, 152
Secondary structure
predict RNA, 314
Secondary structure prediction, 251, 313
Select
exact positions, 127
in sequence, 129
parts of a sequence, 129
workspace, 79
Select annotation, 129
Selection mode in the toolbar, 77
Selection, adjust, 129
Selection, expand, 129
Selection, location on sequence, 77
Separate sequences on gel, 276
using restriction enzymes, 276
Sequence
alignment, 282
analysis, 195
display different information, 67
extract from sequence list, 145
find, 127
information, 141
join, 214
layout, 123
lists, 143
logo, 313
logo Bioinformatics explained, 290

new, 142
region types, 130
search, 127
select, 129
shuffle, 195
statistics, 208
view, 122
view as text, 142
view circular, 130
view format, 67
web info, 158
Sequence logo, 289
Sequencing data, 311
Sequencing primers, 314
Share data, 63, 311
Share Side Panel Settings, 92
Shared BLAST database, 173
Shortcuts, 80
Show
results from a finished process, 78
Show dialogs, 90
Show/hide Toolbox, 79
Shuffle sequence, 195, 312
Side Panel
tutorial, 40
Side Panel Settings
export, 92
import, 92
share with others, 92
Side Panel, location of, 91
Signal peptide, 234, 235, 313
SignalP, 234
Bioinformatics explained, 235
Single base editing
in sequences, 130
Single cutters, 267
SNP detection, 311
Solexa, *see* Illumina Genome Analyzer
SOLiD data, 311
Sort
sequences alphabetically, 293
sequences by similarity, 293
Sort, folders, 65
Source element, 117
Species, display name, 67
Staden, file format, 327
Standard layout, trees, 305
Standard Settings, CLC, 96
Star activity, 278
Start Codon, 230
Start-up problems, 29
Statistics
about sequence, 312
protein, 211
sequence, 208
Status Bar, 78, 79
illustration, 62
.str, file format, 329
Structure scanning, 314
Structure, prediction, 251
Style sheet, preferences, 95
Support mail, 12
Surface probability, 126
.svg-format, export, 111
Swiss-Prot, 152
search, *see* UniProt
Swiss-Prot, file format, 327
Swiss-Prot/TrEMBL, 312
.swp, file format, 329
System requirements, 15

Tab delimited, file format, 329
Tab, file format, 327
Table of fragments, 274
Tabs, use of, 69
Tag-based expression profiling, 311
TaqMan primers, 314
.tar, file format, 329
Tar, file format, 329
Taxonomy
batch edit, 69
tBLASTn, 163
tBLASTx, 162
Terminated processes, 78
Text format, 128
user manual, 35
view sequence, 142
Text, file format, 329
.tif-format, export, 111
Tips for BLAST searches, 47
TMHMM, 241
Toolbar
illustration, 62
preferences, 91
Toolbox, 78, 79
illustration, 62
show/hide, 79

Topology layout, trees, 305
Trace colors, 125
Trace data, 311
Translate
  a selection, 126
  along DNA sequence, 125
  annotation to protein, 129
  CDS, 230
  coding regions, 230
  DNA to RNA, 225
  nucleotide sequence, 228
  ORF, 230
  protein, 255
  RNA to DNA, 226
  to DNA, 313
  to protein, 228, 313
Translation
  of a selection, 126
  show together with DNA sequence, 125
Transmembrane helix prediction, 241, 313
TrEMBL, search, 152
Trim, 311
TSV, file format, 327
Tutorial
  Getting started, 37
.txt, file format, 329

UIPAC codes
  amino acids, 330
Undo limit, 89
Undo/Redo, 72
UniProt, 152
  search, 152, 312
  search sequence in, 159
UPGMA algorithm, 307, 314
Urls, Navigation Area, 109
User defined view settings, 91
User interface, 62

Vector graphics, export, 111
VectorNTI
  file format, 327
View, 69
  alignment, 288
  dot plots, 199
  GenBank format, 142
  preferences, 75
  save changes, 72
  sequence, 122
  sequence as text, 142
View Area, 69
  illustration, 62
View preferences, 90
  show automatically, 91
  style sheet, 95
View settings
  user defined, 91
Virtual gel, 315
.vsf, file format for settings, 92

Web page, import sequence from, 104
Wildcard, append to search, 149, 152, 155
Windows installation, 12
Workspace, 79
  create, 79
  delete, 80
  save, 79
  select, 79
Wrap sequences, 123

.xls, file format, 329
.xlsx, file format, 329
.xml, file format, 329

Zip, file format, 327–329
Zoom, 76
  tutorial, 38
Zoom In, 76
Zoom Out, 76
Zoom to 100% , 77
Zoom, 3D structure, 187